DOCUMENT RESUME

ED 230 558 TM 820 263

AUTHOR	Harris, Chester W.; And Others
TITLE	Studies of Domain Referenced Item Models. Final Report. Volume 2: Sections III and IV.
INSTITUTION	California Univ., Santa Barbara.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC. Teaching and Learning Program.
PUB DATE	15 Jul 80
GRANT	NIE-G-78-0085
NOTE	183p.; For related document, see TM 820 262.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	*Achievement Tests; Classroom Research; *Difficulty Level; Elementary Education; Estimation (Mathematics); Instructional Materials; Latent Trait Theory; *Maximum Likelihood Statistics; *Models; Sampling; *Test Items
IDENTIFIERS	Monte Carlo Studies

ABSTRACT

The third section of this two-volume report examines the utility of the model developed in section I for application in a classroom testing situation. The model was applied in arithmetic classes through a weekly testing program. The teacher specified the generic task being taught, tests were constructed using a table of random numbers to select the numbers for addition or subtraction, and results were reported to the teachers. The results were used in three ways: as information about individual and class performance, for planning further instruction, and for providing feedback to students. A miniature study of subtraction was undertaken to illustrate the use of conventional analysis of variance procedures with these tests. The results of the study suggest that short tests developed without preliminary item analyses and primarily for use in monitoring classroom instruction can also be used in a more conventional fashion. The final section of the report contains an annotated bibliography of related research studies. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

STUDIES OF DOMAIN REFERENCED ITEM MODELS

UNIVERSITY OF CALIFORNIA
SANTA BARBARA, CALIFORNIA

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

NATIONAL INSTITUTE OF EDUCATION
Program on Teaching and Learning

SECTION III

Contents

	Page
The Generic Task	2
Applications to Classroom Testing	4
Test Results	13
Retention over the Summer Interim	14
A Miniature Study of Subtraction	38
Appendices	45

Tables

1.	Retention Study: Addition	39
2.	Retention Study: Subtraction	40
3.	Retention Study: Multiplication	41
4.	Average Subtraction Item Difficulties	42
5.	Anova Results, Subtraction Study	43

SECTION III

In the first section of this report we described our point of view about testing in connection with classroom instruction and we developed the model that we are studying. This model postulates a generic task (such as addition of two-digit numbers) and uses a random sampling principle to generate sample items, the performance on which gives an unbiased estimate of the proportion of items in the universe that can be answered correctly by the student or examinee. This model is restricted to a generic task for which it is unlikely that the examinee can "guess right"; as such it does not apply to multiple-choice items. Instead the model is more closely related to the content standard notion of Ebel which was described in Section I.

The notion of an item universe plays an important role in our study. Let us contrast two approaches to test development that might be chosen --one that emphasizes performance on a particular set of items (the test) and uses various procedures to select an optimum set of items, and another that emphasizes an item universe, the performance on which is to be estimated, and selects what may be called randomly equivalent sets of items as tests. For example, one might construct a number of addition items and call this a "pool"; such a pool is not necessarily the universe of all possible such items. One could then try out a number of items and use one of several procedures to secure for each of the items studied an index or a pair of indices that describe the item in some specified fashion. One then might select items in terms of these indices to make up a test or possibly two or more so-called equivalent tests. A basic feature of this type of test development is that a continuum of examinee ability or achievement is postulated as essentially the criterion against which performance on an item is judged.

Early, and rather simple, item analysis procedures included the correlation of item performance (a biserial or point biserial correlation coefficient) with performance on some group of items that was taken as a surrogate for the postulated "ability" or "achievement" dimension which ordinarily was conceived as a latent and unobservable trait. Such a study would yield for each item a measure of difficulty (proportion right or, more properly, the "easiness" of the item) and a measure of association of the item score with the score given by the sum of the item performances. One could then select items with a desired pattern of difficulties that correlated well with the surrogate, and use these as a "test." Today we have more sophisticated procedures based on

one-parameter, two-parameter, or three-parameter item characteristic curve models. These models also require a postulated continuum of "ability" or "achievement" as a criterion and yield (at least) a measure of how well the item discriminates, which is essentially a measure of how well the item is associated with the surrogate variable. One can then choose items that discriminate well at a certain level or levels of difficulty to build a test, and this test should function well to sort examinees or distinguish among examinees at this point or these points.

In practice, it is recognized that for item analysis or item characteristic curve analysis to be useful a substantial number of examinees must be tested; consequently these procedures probably are most useful for large-scale testing programs--programs of commercial test makers, of state-wide testing such as California's, and the like.

It is at this point that one may ask about the utility of such procedures for the case of a single classroom group of students who are being taught certain specific task performances or problem-solving procedures. First, this number of students may be 20 to 30, and there may not be any other available students who have had a similar instructional history and are being instructed currently in the same fashion. Instruction is designed to change the performance of the examinee and, usually, to change the performance rather rapidly; if this occurs, then examinee achievement is not stable over time but improves. Furthermore, for any given set of students in a classroom, the change in performance probably proceeds at different rates for different students; if so, the achievement continuum that must be postulated for the first approach to test development changes with instruction and furthermore these changes are not simply a consistent displacement of the distribution some constant number of units on the continuum. One then wonders whether data gathered at one time lead to a test that has similar optimum characteristics at a later point in the instruction. This is an especially important question if the instruction is effective and the students are changing rapidly. If the schooling has no effect, of course, then the item analysis or item characteristic curve procedures will develop tests that are very much like conventional academic aptitude or intelligence tests.

Let us now look more closely at classroom instruction. It may be true that academic aptitude or intelligence tests provide useful data for organizing classrooms, possibly in selecting curriculum materials, and the like. However, classroom instruction often is designed to bring about fairly specific changes in performance in the students, and these changes are seldom evidenced in aptitude or intelligence test scores. How then might tests be developed that would give a teacher information about the changes in performance that the teacher hopes will take place?

The generic task. There are a number of school objectives that can be viewed as the improvement of student performance on a generic task. Given an appropriate set of words for the age and instructional level of the students, the spelling of such words is a generic task. Given a base 10 (or for that matter, base 2 or base 12) number system, the addition or the subtraction of numbers in this system is a generic task.

Given appropriate passages, the identification of the main idea or of the topic sentence of such paragraphs is a generic task. For any such generic task, there will exist a number of specific tasks each of which employs a specific content. Thus

$$\begin{array}{r} 29 \\ +14 \\ \hline \end{array}$$

is a specific task in the addition of two two-digit (base 10) whole numbers. The generic task plus the subject matter or content that is specified in the objective define a universe, the performance on which is a measure of the current achievement of a student or of a group of students. Thus at various times in the educational history of a student, we may want to ask how well he can add two whole numbers, how well he can add two or more two-digit numbers, how well he can add several two-digit numbers for which regrouping ("carrying") is required, how well he can add several decimal numbers, and the like. Each such question conceptualizes a universe of specific tasks that is homogeneous in the sense that all the specific tasks are an example of the same generic task.

Thus the generic task plus the content or subject matter define a universe of performances, and these specific tasks can be taken as the foundations of achievement items. In order to prepare such items for use with students it is necessary, in addition, to define the conditions under which the specific performance is to be observed; this definition will in effect specify item format, time of testing, mode (oral or written) of testing, etc. It should be clear that the choice of conditions may be arbitrary. For example, the item

$$\begin{array}{r} 29 \\ +14 \\ \hline \end{array}$$

and the item

$$29 + 14 = \underline{\quad\quad}$$

differ in format and may measure at least slightly different achievements. However, it is not unreasonable to include as part of instruction, information about common formats and about the formats that will be used in exercises and/or tests. The analysis of an instructional objective should describe the generic task, the appropriate subject-matter or content, and the various conditions under which student performance is to be observed. With these spelled out it is then possible to generate items to be administered to the students.

If it were feasible and economical, one might want to observe the performance on every item in the universe; however, this ordinarily cannot be done. For example, there are 10^4 items for the addition of two two-digit (base 10) whole numbers, if we allow zero to appear in any of the four positions. If a student requires 30 seconds to work each problem, then 83-1/3 hours would be required to complete all the 10,000 items.

It is obvious that a testing session this long is impossible for any student. Let us grant that except in possibly trivial cases of small finite item universes, it is never possible in practice to secure a universe score for any student. What then can be done? The item analysis or item characteristic curve approach attempts to solve the problem by identifying a set of items that in effect are highly correlated with the surrogate variable, and to develp estimates of ability from the scores on this set of items. Note that the universe scores still are not available, and that the surrogate variable, depending upon how it is constructed, may or may not be a well behaved estimator of the universe score. Also, note that what is regarded as a high correlation between the score on the items selected for the test and the score on the surrogate still may allow large discrepancies in the rank order of students on the two variables. Simply observe a fairly regular scatter plot with a correlation of, say, .50 to see this.

Another approach is to build tests by sampling at random from the items in the item universe. Interestingly, this can be done even though the entire item universe is not actually written out or specifically available. Again consider the addition generic task we have used before. The 10,000 items of this universe are <u>implicitly</u> available, and a random sample of these items can be constructed readily by randomly selecting digits (including zero) to fill the four spaces for each item. It then follows that a test of m items constructed in this fashion yields a score (number right) that is an unbiased estimate of the (unavailable) student's score on the item universe. In other words, if we secure the performance of a given student on such independently selected m-item tests, the average score for the student approaches the student's average universe score as the number of tests increases. Thus an estimate that is fairly stable can be secured by choosing m to be large enough. We note that for this approach, a test itself is not of primary concern; our concern is with the universe score of which the test yields an estimate.

<u>Applications to classroom testing</u>. We took this notion about the construction of tests that are related to classroom instruction, and our ideas about the estimation of the parameters of our item model into classrooms, administered tests over a period of time, and worked with the teachers on the interpretation of the data yielded by these classroom tests.

Our first task was to identify some teachers who were carrying out a program of arithmetic instruction whose objectives would permit the type of test construction we were studying. We wanted the teachers to be able to specify on a week-to-week basis the generic task or tasks they were trying to teach; we then showed the teacher examples of items and item formats fitting this specification and the teacher chose the type of item that the teacher regarded as appropriate. We also suggested a test length and a time at which we would come to the class, administer the test or tests, score the tests, and give the teacher the results. In general we attempted to carry out weekly testing over a period of time, with the understanding that the generic task being tested could change from week to week in keeping with the teacher's instructional plans and the progress of the students. We also provided for randomly

equivalent tests to be used two or three weeks in a row when the teacher judged that the students needed further instruction and practice on a particular type of task. As the data presented later in this section indicate, we employed several equivalent forms for a given task over a period of several weeks in some instances. We also used quite a few different task types in keeping with the teacher's instructional program.

Perhaps the most important feature of this testing program is that it was tailored to the instruction being planned for a particular classroom and a particular group of students. Another important feature is that it provides evidence over time of the performance of the same group of students. A third feature is that relatively small (20 to 30) groups of students were the focus of any particular test. Thus this testing experiment differs markedly from typical uses of standardized tests in school-wide, district-wide, or state-wide testing programs. We believe that tailoring the test to what the teacher actually is attempting to teach at that time and testing the same students over time to determine progress give this testing program a basis in instruction that is unique.

The test construction process was carried out by us as soon as we had, for a particular group of students, the teacher's specification of the generic task (or tasks) and of the desired item format. We did restrict item formats to those for which the student was required to produce the response (rather than a multiple-choice item format); this was dictated by the model we have described earlier which assumes that the examinee cannot "guess right." To build the items required for a test, we used a table of random numbers to select digits (including zero) for each of the positions in the item form. For example, there are nine positions to be filled for any addition problem with three addends, each consisting of three whole numbers. For a division problem with a two-digit whole number divisor and a three-digit whole number dividend, there are five positions to be filled. By selecting digits at random from a random number table to form an item, we in effect created with a set of these items a random sample of all the possible items of that type. We also were able to introduce conditions which in effect defined an item universe that is a sub-universe of a larger one. For example consider the subtraction of a two-digit whole number from a three-digit whole number. Some of these items require no regrouping ("borrowing") and can be solved correctly simply by writing down the differences for two pairs of numbers, as in this example:

$$\begin{array}{r} 857 \\ -34 \\ \hline \end{array}$$

Our procedures permitted us to sample randomly from this sub-universe of items without regrouping, as well as from the entire universe. We illustrate this below in a miniature study of "borrowing" in subtraction.

The tests that were constructed were typed on a master, using an IBM "Orator" ball which prints large numbers. Their size can be observed in the examples attached to this section.

We then had the test duplicated on (usually) colored stock; this was useful whenever we had two tests to be administered to the same group of students during the same period, as we often did. Generally we could place 10 items on a sheet; consequently if the teacher wished a test of 20 items on a particular occasion we created two randomly equivalent 10-item tests, on sheets of different color, and we could administer them separately, using the color code to identify the test. At other times a teacher asked for two different tests on the same occasion, such as an addition test and a subtraction test; again the color code simplified administration and the scoring the tests. We scored the tests as soon as they were completed by the students by marking items as correct or incorrect and recording the number correct on the top of the test sheet beside the student's name. These marked tests were given to the teachers who used the data as they saw fit.

We also recorded the results for our records, recording 1 or 0 (correct or incorrect) for each item for each student. We then subsequently entered our data into the local computer and secured estimates, for that group of students on that type of test on that occasion, of the parameters of our item model. These parameters are k, the proportion of students in the + category of the dichotomous latent class variable, and x_t the conditional error rate for each of the items in the test. We interpreted the estimate of k as an estimate of the proportion of a population of students (similar to the classroom sample in instructional history) who understand and can perform the generic task, i.e., belong to the + latent class. Our expectation was that the value of the estimated k would increase beginning with the introduction of the skill being taught through the period of systematic instruction and practice.

Different teachers used the test results differently. Three major uses appeared to evolve: (1) providing information to the teacher about individual and class performance on tasks emphasized during the previous week's instruction; (2) providing information to the teacher for planning further instruction; (3) providing immediate feedback to students about their errors and offering reinforcement for work well done. It was our belief that having the tests prepared, administered, and scored by "outsiders" (i.e., by us) helped to establish a serious working atmosphere in the classroom. Several times we observed this effect of our entrance into the classroom when a substitute teacher was in charge that day; when we passed out the test papers and began the testing, the students became serious about their work.

Teachers often used the test results to discover the specific problems students had with the previous week's material, and these observations often influenced the plans for future instruction. For example, if students did poorly on a particular set of problems, the teacher might spend more time on this area during the next week and ask us to prepare a similar set of items for the next week's test. If the students did well, the teacher would likely move on to a new area and ask us to prepare items over this new domain for the coming week. Also, there were times when teachers requested new tests over tasks that had been covered earlier in the year as a way of determining how well students had retained the material.

Many times teachers returned the scored test papers immediately to the students so that they could discover and correct their errors; thus students received immediate feedback that showed them how well or poorly they had done and had an opportunity to redo the problems they had missed. Since the tests were relatively short, students could work a set of problems, have them scored by us, and correct their errors all within the time period allocated to arithmetic for that day. Teachers also often charted the scores from week to week for individuals and for the class as a whole, and some teachers posted this information on bulletin boards.

Our impression was that the teachers who participated in our program believed that they were helped in several ways. For one, we constructed the tests and relieved the teacher of that burden. For another, the presence of outsiders for a short period of time once a week helped to change the classroom atmosphere and emphasize the seriousness of the tests. The immediate scoring of the tests which permitted prompt feedback to teachers and to students was highly regarded. (Note that school-wide or district-wide testing programs usually involve a long delay between test administration and the communication of information about the performances.) Also the weekly program gave regular practice in test-taking spaced over the school term; this regular practice may have helped students to retain what they were learning.

ADDITION

NAME _____

8823 133 27 + 6	635 54 9 +5788	18 6 372 +6955
8 207 1761 + 28	45 6306 749 + 7	2 76 559 +8148
599 2 70 +4285	3824 41 983 + 5	
9258 3 37 + 684	380 76 4 +7885	

SUBTRACTION

NAME _____

549 -328	870 -417	675 -524
764 -248	986 -753	429 -306
851 -635	297 -138	
386 -253	645 -416	

MULTIPLICATION

NAME _____

132 × 8	479 × 6	865 × 2
735 × 4	967 × 3	791 × 5
524 × 6	387 × 8	
297 × 7	904 × 9	

ADDITION

NAME _____

730 17 325 + 23	41 961 444 + 37	197 25 87 +466
923 43 98 +820	87 82 469 +209	168 34 341 + 91
65 72 333 +860	526 902 68 + 29	
95 335 501 + 91	22 723 119 + 90	

MULTIPLICATION

NAME _____

134 x427	866 x245	349 x501
749 x687	848 x542	739 x234
397 x820	170 x957	
867 x511	300 x959	

Test results. We present in the Appendices for Section III tables of data derived from classroom testing over a period of approximately one and one-half years. In each table we identify the arithmetic task and its specific item format, and then give information for the various classes who took this type of test. The information includes the date of testing. As these dates indicate, quite often the same generic task (but different sets of items) was administered to the same class on several occasions. We used the Goodman procedure, described in Section II, to estimate k, the proportion of examinees who understood the generic task, and x_t or the specific item difficulty for each item. The range of $\hat{x}_t$ and the average $\hat{x}_t$ are presented for each administration. We also present the number of examinees (n) and the number of such items (m). Thus if one accumulates the product of n and m over all the tables in the appendices, one can determine how many bits of data (item-examinee responses) were collected. This number is quite large.

The appendices present the data for the four arithmetic tasks separately, beginning with addition and ending with division. Certain observations are prompted by the data.

For one, it is evident that different teachers were teaching the same type of task. Recall that any test we constructed was specifically requested by a teacher and the item format approved by the teacher. It certainly isn't surprising that elementary school teachers of students in grades 3 to 6 were teaching arithmetic over a period of time, nor that they emphasized the four "fundamental" operations. Our data indicate that different teachers often regarded the same generic task as an appropriate one to emphasize.

A second observation is that within a class the same generic task often was tested more than once, with testing sessions commonly held a week apart. This aspect of the record indicates that teachers often continued the testing (and of course the teaching) of the same task over a period of time in an effort to bring the class to an acceptable level of performance. The value of the average $\hat{x}_t$ probably is a good index of performance (especially when the value of $\hat{k}$ is large and fairly consistent) over time. As can be seen from some of the tables this repeated testing of the same task (with different items) was accompanied by decreases in the average $\hat{x}_t$, an indication that the items were being answered correctly more often. For other parts of the record there is a clear indication that $\hat{k}$ increases over the repeated testings of the class; this suggests that understanding of the generic task is increasing with the testing and instruction.

A third observation is that the overall difficulty, indexed by $\hat{k}(1 - \text{average } \hat{x}_t)$, differs for the four arithmetic operations among these grade 3 to grade 6 students. In general addition is easier, which is not surprising, and multiplication and division are more difficult. As the dates of testing indicate, the teachers usually emphasized addition and subtraction earlier in the year and multiplication and division later. This too reflects a rather conventional view of the arithmetic curriculum.

One might also use these data to identify differences between classes. It was not our intent to use the testing procedure in this way, and we did not (with two exceptions) use exactly the same tests with more than one class. (These exceptions are described below in the reports of the retention and the subtraction studies.) Instead, each test was constructed for use with a particular class at a particular point in its instructional history and was designed to meet a particular teacher's criterion of appropriateness. Thus comparisons of groups, although they may be made, are not planned for and may be somewhat misleading. In general we would recommend that these test data be examined for evidence of improving achievement on the part of a group rather than for evidence of different levels of achievement among groups.

Finally we found that the Goodman method of estimating k and x_t functioned very well for these data, even though we usually had small examinee groups of 20 to 30 and short tests of 5 to 10 items. The estimates converged and converged rapidly. We regard the Goodman method highly.

Retention over the summer interim. An important question is that of retention of what has been learned from classroom instruction. In order to throw some light on this we abstracted a number of research studies that examined this question. These abstracts follow. Note that many of these studies deal with reading. Only a few (see, e.g., Kurtz, R.) provide information about retention of specific arithmetic skills. Further, many of the studies used standardized achievement tests which usually are not focused on specific skills.

Begle, E.G., Ginther, J.R., Pence, B., and Davis, M. Mathematical retention over the summer. Teacher Corps Mathematics Work/Study Group. Working paper No. 1. (ED 138547).

The purpose of this study was to investigate mathematical retention of Junior High School students over the summer months. Approximately 41 % of the eighth grade sugjects used in this study were chicano; the others were anglo, black, oriental, and American Indian. The subjects were divided into four groups (sample size ranged from 25 to 40) according to ethnic origin and according to whether or not they had received pretraining on the tests. The retention tests administered in June (seventh graders) and again in September (eighth graders from the same group), measured (1) mathematical reasoning, (2) computation, (3) comprehension, (4) the ability to read mathematical prose.

Sixteen means and standard deviations were derived for each group for each of the four tests in June and again in September. Fourteen reliabilities were also computed in June and again in September. The reliabilities for two groups were not computed because they were predominantly anglo students, and the Missing Words Test had been found to be reliable for all anglo groups.

The authors concluded that in general, no loss of comprehension or mathematical reasoning over the summer months was found, and the ability to read mathematical prose seemed to increase slightly over the summer months.

Brueckner, L.J., & Distad, H.W. The effect of the summer vacation on the reading ability of first-grade children. The Elementary School Journal, 1924, 24, 698-707.

The authors examined the reading retention ability of students in twelve first-grade classrooms. Using the Minneapolis Primary Reading Test and the Haggerty Reading Examination, Sigma I, they found that median scores for each grade were lower in September than they were in June on the former test, but on the Haggerty Reading Examination there was no difference.

Cohen, A.D. The Culver City Spanish immersion program: How does summer recess affect Spanish speaking ability? Language Learning, 1974, 24, 55-68.

This study looks at one aspect of second-language mastery in depth: patterns of foreign language retention among youg children after being removed from a language contact situation for a period of time. The subjects were 14 Anglo children from the Culver City Spanish Immersion Program, a pioneering project in American public school education. These children were immersed exclusively in Spanish during their kindergarten year. English was gradually introduced in first grade. This report deals with the effects of summer recess between first and second grade upon the spoken Spanish of the students. They were given an Oral Language Achievement Measure individually on a test-retest basis. The results showed that a summer recess of three months took its toll on Anglo children's performance in Spanish. Utterances became shorter; at least one grammatical class (prepositions) was used slightly less while another (verbs) became more prominent; the children made more errors proportionate to what they said; problems with article/adjective agreement not only persisted, but in the case of the definite article, shifted in nature; the ser verb began to be used more than estar when children were in doubt; and inflection for person in present tense indicative verbs continued to cause minor problems.

David, J.L., & Pelavin, S.H. Evaluating compensatory education: over what period of time should achievement be measured? Journal of Educational Measurement, 1978, 15, 91-99.

The authors point out that evaluations of compensatory education programs, in general, have not included measures of sustained achievement. Instead, judgments of program succes have been based on students' achievement during the school year: that is, on a spring posttest score adjusted in some way for the preceding fall pretest score. They hypothesized that evaluations based on measures of sustained achievement would lead to different conclusions than evaluations based on fall-to-spring achievement. Specifically, they expected that evaluations based on a fall-to-fall period, by virtue of including the summer months, would result in smaller achievement gains than traditional fall-to-spring evaluations.

The authors obtained longitudinal data from evaluations of several compensatory-education programs. They used data from those programs that administered standardized achievement tests annually in both the fall and spring for consecutive grades. The analysis consisted of comparing achievement over the 7-month fall-to-spring period with achievement over the 12-month fall-to-fall period. They did this by calculating the means for each test point for all samples with three test administrations (fall, spring, fall) and then subtracting pretest from posttest to obtain estimates of mean gain for the two time periods.

The authors maintain that the results of the data support their hypothesis that the inclusion of the summer months in the evaluation period will result in smaller gains in achievement: in all five samples, the second fall score is smaller than the spring score. They conclude that there can be large and statistically significant achievement losses over the summer and, hence, fall-to-fall gains can be considerably smaller than fall-to-spring gains. They point out that if evaluations of compensatory-education programs were conducted on a fall-to-fall basis, conclusions about program success might be quite different.

Elder, H.E The effect of the summer vacation on silent-reading ability in the intermediate grades. The Elementary School Journal, 1927, 27, 541-546.

Elder tested 203 subjects in May and September with the Monroe Standardized Silent Reading Test. When the data were analyzed without reference to grade, it was discovered that 59% of the subjects improved, 27% regressed, and 15% remained the same. The average gain per pupil during the four and a half month interval was .45 of a school grade. Elder also found the range of performance to be greater in September than in May. However, an important limitation in the Elder study was the fact that the May test was administered almost one month before school adjourned for the summer.

Haydon, J.R., Davis, D., Bowman, J, Pritchard, J. Evaluation report, Title I ESEA Project activities, 1966-67. (ED 016744).

Each of the 22 compensatory education projetts conducted in the Des Moines, Iowa, public schools during the school year and summer was separately evaluated in this report. Some of the projects concentrated on instruction in reading, language arts, mathematics, humanities, or practical science. Others offered special social and health services. A few projects provided enrichment, tutoring, instruction in small classes, or therapy for handicapped children. The number of students showing academic gain or loss on standard achievement tests is indicated in the report, but the projects' evaluation techniques and findings are not discussed.

Hillerich, R.L. Pre-reading skills in kindergarten: a second report. The Elementary School Journal, 1965, 65, 312-317.

The effect of summer vacation on the retention ability of kindergarten subjects exposed to two different programs was the focus of Hillerich's investigation. He examined subjects' ability to use context, find letters, listen for letter sounds, and match letters and sounds. Subjects enrolled in kindergartens where a workbook was used retained significantly ($p < .01$) less than non-workbook subjects. The mean loss on the fifty-eight items was only 2.15 raw score points, however. The author stated that this was evidence that the skills were retained over the summer vacation.

Irmina, Sister M. The effect of summer vacation on the retention of the elementary school subject. Catholic University American Research Bulletin, 1928, pp. 3-99.

The author examined retention ability of first through seventh grade subjects using eleven different measures, including intelligence, reading, math and spelling ability. Tests were administered the last week of school in the spring and within two weeks after students returned to school in the fall. She concluded that the word recognition ability of first and second graders was not seriously affected by the vacation period. The word reading, phrase reading, and sentence reading subtest scores loss of first grade subjects was significant, but for second graders the loss was slight in two of the schools studied and a gain was reported in the third school. On the Reading of Directions subtest of the Gates Primary Reading Tests, a consistent loss was found in grade one. In grade two, however, the subjects in two of the schools reported a slight gain. Subjects in all three schools indicated a gain in reading ability at the second grade level as measured by the Stanford Primary Reading Examination, Paragraph Reading subtest. The author concluded that there appeared to be no actual loss in reading ability due to a non-school period.

Keyes, N., & Lawson, J.V. Summer versus winter gains in reading school achievement. School and Society, 1937, 46, 541-544.

The authors tried to determine the stability of subjects' standardized test scores over an extended period of time. Their subjects originally included 164 4th, 5th and 6th graders in Gilbert Minn. Subjects were tested on the Unit Scales of Attainment each fall and spring between 1933 and 1937. The test included eleven subtests: reading, arithmetic operations, problem solving, American history, geography, elementary science, literature, spelling, english usage, capitalization, and punctuation. Tests were administered one month before school dismissed in Spring and were readministered one month after school resumed in the fall. The investigators found that the reading score did not decrease during the five-month interval. There was a loss in the mean arithmetic, science and literature scores, however.

Kurtz, R. Fourth-grade division: how much is retained in grade five. The Arithmetic Teacher, 1973, 20, 65-71.

The design of the study provided for the testing of all fourth graders in a small city school system in Kansas and the retesting of those same children at the beginning of the fifth grade. Only scores from the 343 students (163 boys and 180 girls) who were able to participate in both fourth- and fifth-grade testings were included in the data analysis. It was hypothesized that the results of the pretests and posttests, which were administered in May and September respectively, would show that over the summer fifth graders lose a significant amount of the division skills that they possessed at the end of the previous year.

A 16-item test was devised to assess various computational skills in division at the fourth-grade level. The items covered a wide range of difficulty from problems with a single digit divisor and a single digit dividend without a remainder, to a two digit divisor and a two digit dividend with a remainder.

The overall-performance comparisons of the study show that, on the average, fifth-grade students, after the lapse of summer, were able to work approximately two less problems than they were able to work at the end of the fourth grade. The boys recorded virtually the same loss as did the girls. While indications are that losses over the summer were great, the relatively low (63%) accuracy level achieved by the end-of-year fourth graders indicates that the skills were not well learned.

After comparing the summer loss for students in the various quartiles, a pattern emerged suggesting that students in the upper quartile registered considerably more summer loss than students in the lower quartile. This information supports the position that even the best fourth-grade students need considerable review in the fifth grade. The fact that fourth-grade students who scored in the

lowest quartile were able to work little more than 25 % of the problems correctly in the fifth grade verifies the need for developmental instruction in the division processes. The author believes that a short review might be sufficient for upper quartile students, but it is inadequate for those in the lower quartile.

A survey of the most difficult problems for fourth graders showed that the same problems were the most troublesome for fifth graders. However, there was little loss in the ability to work these problems. Kurtz quotes the old phrase, "you can't forget what you never learned".

The author concludes that this study presents sufficient evidence to support the conclusion that mastery of fourth-grade division is not satisfactorily attained by over one-half of the fourth graders. When this mastery level is further reduced by the attrition caused by summer vacation, it is evident that a fifth-grade teacher should be prepared for considerable variation in division skills in each new class.

Montgomery, J.L. A comparison of BSCS versus traditional teaching methods by testing student achievement and retention of biology concepts. Report submitted to the East Central Indiana Curriculum Improvement Project. Muncie, Indiana: Ball State University, May, 1969. (ED 033866).

The purpose of this study was to examine the effect of the biological sciences curriculum study (BSCS) materials and the inquiry teaching method on student achievement and retention in biology. Teachers were selected who used BSCS materials with inquiry methods, BSCS materials with traditional methods, traditional materials with inquiry methods, and traditional materials with traditional methods. Twelve students selected at random from the classes of each teacher chosen were pre-tested and post-tested using the Nelson Biology Test and the Processes of Science Test as a measure of achievement; the same tests were administered after the summer vacation as a measure of retention. The data were analyzed by analysis of covariance using the pretest scores as covariates in analyzing the post-test scores, and the post-test scores as covariates in analyzing the retention scores. The results indicated that the BSCS students taught by inquiry methods showed the greatest achievement, all BSCS students showed greater retention, inquiry taught traditional students showed greater retention on the Processes of Science Test than traditionally taught traditional students, and tenth grade students out-performed nineth grade students. There was also a positive relationship between class size and both achievement and retention.

Morrison, J.C. What effect has the summer vacation on children's learning and ability to learn? Ohio State University Educational Research Bulletin, 1924, 245-249.

Using the Haggerty Reading Examination, Sigma I, Morrison found that the median scores of first grade subjects actually increased over the summer. However, his sirst grade sample included only forty-five subjects, all from the same school. When scores from subjects in grades one, two, and three were examined together, 70% of the subjects improved while 30% lost. He concluded, however, that there was practically no change in their reading ability.

Parsley, K.M., & Powell, M. Achievement gains or losses during the academic year and over the summer vacation period: A study of trends in achievement by sex and grade level among students of average intelligence. Genetic Psychology Monographs, 1962, 66, 285-342.

The authors randomly selected ninety males and ninety females at the second through seventh grade levels with intelligence quotients between 90 and 110 as measured by the California Test of Mental Maturity. The California Achievement Test was used to determine if reading ability was retained over the summer vacation period. They found that reading vocabulary scores tended to increase, by grade level, over the summer, up to grade five. Or, while there was a slight loss in mean scores at the second grade level, there was no loss at the third grade level and an actual gain at the fourth and fifth grade level. A similar trend was evident when reading comprehension socres were examined across the grades.

Perez, Samuel A. The effects of the summer vacation on reading retention. Paper presented at the annual meeting of the International Reading Association, Houston, Texas, 1978. (ED 158243).

To assess the effect of summer vacation on the overall reading ability of first through fifth graders, as measured by norm-referenced and criterion-referenced reading tests, a study was conducted involving 84 children enrolled in the Wisconsin Design for Reading Skill Development-Word Attack at the Edith Bowen Laboratory School in Logan, Utah. More specifically, the question being asked was: What are the specific reading skills that are more easily retained and the specific reading skills that are more difficult to retain by children receiving reading instruction in an objective-based reading program?

In the spring testing the researcher administered the Gates-MacGinitie Reading Test, Form 1 (considered to be a norm referenced test), and the Wisconsin Tests of Reading Skill Development, Form 1 (considered to be a criterion-referenced test) the last two weeks prior to summer vacation. In the Fall testing, Form 2 of the same tests was given. A summer activities questionaire was administered to exclude children who had received reading instruction during the summer.

Three hypotheses were formulated: (1) There is no difference between Spring and Fall mean scores of the Gates-Macginitie Reading Test; (2) There is no difference between Spring and Fall variance scores on the Gates-Macginitie Test; (3) There is no difference in the Spring and Fall mean scores measured by the Wisconsin Tests of Reading Skill Development. An analysis of variance was used to determine the degree of reading retention. Hypotheses (1) and (2) were accepted and (3) was accepted with the following qualifications. Statistically significant gains were found on two of the 33 CRT's measuring specific reading ability. The statistically significant gains were recorded on the tests measuring short vowels and middle vowels.

The results suggest that no significant loss in reading ability occurs over the summer vacation; therefore, the author suggests that teachers using objective-based reading programs need not conduct massive retesting or reteaching in the fall.

Rude, R.T. Sex, intelligence, and school reading curriculum as factors influencing summer retention of overall reading ability and specific reading skills of first-grade subjects. Technical Report Number 263. Wisconsin University, Madison: Research and Development Center for Cognitive Learning, 1973. (ED 095515).

This study was designed to assess the effect the summer vacation period has on the reading ability of first-grade subjects, as measured by norm-and criterion-referenced reading tests. The data were analyzed to determine if sex of subject, IQ, or type of school reading curriculum were related to the ability to retain overall reading ability or specific reading skills.

Subjects in the study were 311 first-grade pupils enrolled in nine northeastern Wisconsin elementary schools. Approximately one-half of the subjects were enrolled in an objectives-based reading program while the remaining subjects were enrolled in basal reader curricula.

All subjects were administered the Gates-MacGinitie Reading Test, Primary A, and the Wisconsin Tests of Reading Skill Development-Word Attack, Level B, two weeks prior to and two weeks after the summer vacation period. In addition, the California Short-Form Test of Mental Maturity was administered to all subjects during the spring testing sessions. Subjects with IQ scores which fell within the third or seventh stanines were not included in the data analysis. A multiple analysis of variance statistical treatment was used to analyze the data. Retention of reading scores between the spring and fall was the dependent variable; sex of subject, intelligence, and type of school reading curriculum were the independent variables.

Statistically significant differences were found between the mean spring and fall test scores on eleven of the fourteen measures. Sex of subject and type of school reading curriculum were not significantly related to ability to retain reading skills. Intelligence of subjects was found to be related to retention ability

Rude,R.T. continued

on only two of the measures.

Fifteen percent of the subjects changed from being considered "masters" of the specific reading skills in the spring to being classified as "nonmasters" in the fall. Achieving a score of eighty percent or better on any of the specific skill tests was the criterion for mastery.

It was concluded that even though statistically significant losses occurred on most of the tests, the most meaningful measure of change was the difference between the percentage of subjects considered to have mastered the skills in the spring versus the percentage in the fall. The fifteen percent change between the two times was not considered great enough to suggest massive schoolwide retesting of all subjects in criterion-referenced reading programs. Instead, retesting of subjects might be done on the basis of teacher subjective judgment, thereby reducing considerably, the cost and time necessary to implement such a reading program.

In conclusion, then, sex of subject, intellectual ability, and type of school reading curriculum do not appear to be important variables related to the retention of overall reading ability and specific reading skills. While significant losses were found on eleven of the fourteen measures, when the data were examined in terms of percentage of subjects considered to have mastered the skills in the spring and fall, only fifteen percent of the subjects needed to be recategorized.

Rude, R.T., Foxflover, P., & Niquette, S. Retention of visual and auditory discrimination reading skills. Journal of Education Research, 1975, 68, 192-196.

The authors concluded that kindergarten children tend to retain visual discrimination skills during the summer months while they tend to show a slight loss in auditory discrimination skills.

Scott, L.F. Summer loss in modern and traditional elementary school mathematics programs. California Journal of Educational Research, 1967, 18, 145-152.

Scott reported two studies investigating retention of mathematics ability after receiving instruction in either a modern or traditional mathematics program. Using an analysis of variance technique, he found no significant differences in retention ability attributable to the two instructional programs at the first and second grades in one study. In the other investigation, no significant differences in retention ability were found at the third, fourth, and sixth grade levels. He found a significant difference ($p < .01$) favoring the traditional group in the first study, while in the second study he found a significant difference ($p < .01$) favoring the modern mathematics group at the fifth grade level. Overall, there was little difference in retention between the two programs.

Townsend, A. Growth of independent-school pupils in achievement on the Stanford Achievement Test, Educational Records Bureau, 1951, 56, 61-67.

Using the Stanford Achievement Tests, and following 56 subjects over three grades and two summers, she found the lowest test-retest correlation on the Reading subtest to be .883. Even though high correlations between test scores were found, she cautioned that care must be exercised when interpreting her data. Correlations between two consecutive fall test scores, for example, were usually higher than consecutive spring-fall test scores. It is important to note that in her study, classroom instruction continued one month after the spring tests were administered and, in the fall, another month of instruction was carried out before the post assessment was conducted.

Turner, E.W. The effect of long summer holidays on children's literacy. Educational Research, 1972, 14, 182-186.

The summer holiday was used as a device for holding steady the school-based variable in order to investigate a possible connection between pupils' home/neighborhood backgrounds and their ability to retain reading skills.

A small pilot scheme using a language-based battery of tests to assess a small number of children (N=30) from disparate backgrounds yielded significant results: a survey was then carried out with 83 from municipal housing and 143 children from owner-occupied property. (This study seems to have been conducted in a Middle School in England; the information is scanty).

The author states that an analysis of the results indicated a correlation between background and retention of reading ability. He concludes that the more able children are less affected by what he terms "adverse forces", and the less able seem particularly vulnerable.

Weinberger, J. Temporal retention study on IPI mathematics. Report to U.S. Department of Health, Education & Welfare, Office of Education, 1969. (ED 036181).

In the individually prescribed instruction (IPI) system, once a pupil begins working through the objectives, it should not be necessary to have him take placement tests each fall. The purpose of this study was to determine whether or not it is necessary to give placement tests to the pupils at the beginning of each school year. To meet this purpose, the number of units the pupils have gained or lost over the summer has been calculated by area in the continuum and grade level of the pupil. The data used was extracted from the pupils' placement profiles forthe spring and fall of 1968 in four schools for 1,231 pupils representing grades 1-5. These data were analyzed to determine if the IPI policy regarding a fall placement test was correct. The results of the study show that it is unnecessary to have the placement test again in the fall.

Womble, M.L. Summer recess: Does it make a difference on Title I student achievement? (ED 141445).

A random sample of fourth and eighth grade title I students who either did or did not attend summer school was tested to determine what effect the title I summer school program had on student achievement and summer loss in reading and mathematics. Academic achievement was assessed by the Stanford Achievement Tests. An analysis of covariance was computed on the pretest and posttest scores for both groups of students (summer school and nonsummer school) for each service (reading and mathematics) for each grade level. The pretest scores served as the covariate and the posttest the criteria. Most differences between the two groups were nonexistent by the end of september. Students not attending summer school usually gained more or lost less than students attending summer school. The necessity of a summer Title I academic program was considered questionable.

These abstracts strongly suggest that we do not have adequate and relevant data to determine to what extent classroom instruction in specific arithmetic skills is retained over a period of time such as the summer. Studies that use a testing system such as the one we are studying and compare performance over time for students who actually were given direct instruction and practice on specific skills clearly are needed.

We were able to retest (using the same sets of items) in September 1979 two groups of students who had been tested earlier in that year or later in the preceding year. We present these results in Tables 1 through 3.

We note that for the addition task tested, Class 4 tended to perform in much the same fashion over the period of approximately a year. Class 5 performance may indicate that the specific number combinations have not been well retained; this is indicated by the average $\hat{x}_t$ values. Also note that we were unable to retest 6 of the 25 students in Class 5. For subtraction the results were quite similar on the two occasions for Class 4. Class 5 may have made some progress, if the value of the average $\hat{x}_t$ is examined; however, k decreased slightly. For multiplication the value of the average $\hat{x}_t$ did not change much for either class; however $\hat{k}$ did.

These three sets of results may be most valuable in suggesting a slightly different approach to the study of retention. We believe that for generic tasks such as these, the values of $\hat{x}_t$ are an indicator of the difficulty of the specific number combinations in the item, and the values of $\hat{k}$ an indicator of the proportion of students who understand the generic task. These data suggest that there is no strong evidence across these time periods of either marked growth or marked deterioration with respect to either of these aspects of achievement.

A miniature study of subtraction. Finally we present a small study as an illustration of the use of conventional analysis of variance procedures with our types of tests. For Classes 4 and 5 we prepared 20 subtraction items (randomly placed on two 10-item sheets) of 4 types. All of the items had this format:

$$\begin{array}{r} zzz \\ -zzz \\ \hline \end{array}$$

Five of the items required no regrouping ("borrowing") and could be solved correctly by simply writing down the difference between each pair of numbers. We had observed earlier that some students made a mistake of this type on problems that did require regrouping. The second type required a regrouping in the ten's place only ("borrowing" a ten). The third type required regrouping only in the hundreds place, and the fourth type required regrouping in both the ten's and hundred's places.

Both tests were given to 25 students in each of the two classes on the same morning (but at different hours). The data were then examined for evidence of differences in difficulty of the four types of items for the two classes. Our expectation was that the two classes would perform

Table 1

Retention Study: Addition

zzz
+zz

Class 4					
Date	$\hat{k}$	Range of $\hat{x}_t$	Average $\hat{x}_t$	n	m
10-11-78	.96	.08-.16	.10	26	5
9-27-79	.96	0-.09	.05	24	5
Class 5					
Date	$\hat{k}$	Range of $\hat{x}_t$	Average $\hat{x}_t$	n	m
11- 1-78	.92	0-.17	.10	25	5
9-27-79	.95	.11-.28	.23	19	5

Table 2

Retention Study: Subtraction

$$\begin{array}{r} zzz \\ -zz \\ \hline \end{array}$$

Class 4					
Date	$\hat{k}$	Range of $\hat{x}_t$	Average $\hat{x}_t$	n	m
10-11-78	.96	.08-.28	.20	26	5
9-27-79	.92	.09-.27	.16	24	5
Class 5					
Date	$\hat{k}$	Range of $\hat{x}_t$	Average $\hat{x}_t$	n	m
11- 1-78	.88	.05-.86	.38	25	5
9-27-79	.84	.19-.44	.30	29	5

Table 3

Retention Study: Multiplication

zz
×z

Class 5					
Date	$\hat{k}$	Range of $\hat{x}_t$	Average $\hat{x}_t$	n	m
3-29-79	.88	.18-.59	.37	25	10
9-27-79	.95	0-.78	.41	19	10

zzz
×zz

Class 4					
Date	$\hat{k}$	Range of $\hat{x}_t$	Average $\hat{x}_t$	n	m
4-19-79	.97	.07-.48	.31	30	10
9-27-79	.88	.14-.52	.32	24	10

at a different overall level, with Class 5 more accomplished on the average than Class 4. We also expected that the first type of item (no "borrowing") would be easier than the other three types, with these three types possibly progressively more difficult.

Table 4 presents the average item difficulty value for the eight calls of this design (2 classes by 4 item types). Each such entry is based on 125 responses (25 students by 5 items). These data suggest that the two classes perform at different levels as was expected and that Item-Type 1 (no "borrowing") is easier than the other three item types, also as expected. There is no strong evidence here that item types 2, 3, and 4 are systematically different in difficulty.

Table 4

Average Subtraction Item Difficulties for 2 Classes and 4 Item Types

	Item Type			
Class	1	2	3	4
4	.94	.78	.70	.74
5	.98	.87	.90	.88

These results are summarized in Table 5 which presents the conventional anova table for these data. If we make certain distributional assumptions then we can identify statistically significant effects in this study. The effect (mean square) due to item types, when compared with the variation within types yields an F ratio of 29.40 that is significant well beyond the 1% level, even though we have only 3 and 16 df for this comparison. For classes we have an F ratio of 3.95 which is not significant at the 5% level but is at the 10% level. (Incidentally, this comparison is essentially a two independent group t-test for the examinees' total scores on the 20 items. As such it is the most robust comparison in this study.) The comparison of classes by item types with the source attributable to individuals within classes by item types yields as F ratio of 1.62 which is not significant at the 10% level. This comparison essentially asks whether or not the two classes differ in their "profile" over item types and is an approximation to a multivariate test. No other comparisons seem to be meaningful in this study.

The results suggest that the two classes may differ in overall performance, but this is not strongly indicated. The results also suggest that the item types present different levels of difficulty to the students, a finding that is consistent with much informed teacher experience; here, however, we have isolated a specific question and allowed systematically collected data to reflect on this belief or hunch. Finally, the results do not support the notion that there is an "interaction" of class and item type or a difference in profile over item-type for the two classes.

Table 5

Anova Results, Subtraction Study

Source	df	SS	MS
Items	19	-	-
Between Item Types	3	440.00	1.47
Within Item Types	16	0.78	0.05
Individuals	49	-	-
Between Classes	1	3.36	3.36
Within Classes	48	40.84	0.85
Items by Individuals	931	-	-
Classes by 4 Item Types	3	0.78	0.26
Classes by Items/Types	16	1.14	0.07
Individuals/Classes by 4 Item Types	144	22.52	0.16
Individuals/Classes by Items/Types	768	53.68	0.07

We wish to emphasize the point that this illustrates the conventional treatment of data from short tests devised according to our model. Each of the four types formed only a 5-item test, yet these were adequate to give a dependable answer to the question of differences in task difficulty. It seems likely that these short tests are not highly reliable in a conventional sense, and that this is associated with a lack of power in the comparison of the two classes. We should remember, however, that these tests were not built to maximize the separation of groups by "pretesting" items and selecting them according to difficulty level and discrimination; instead each of these five item tests consists of five exemplars of a generic task drawn randomly from the universe of all such exemplars. We believe that we should be encouraged to find that such short tests developed without preliminary item analyses and primarily for use in monitoring classroom instruction can also be used in a more conventional fashion.

APPENDICES TO

SECTION III

ADDITION

zz
+zz

Class 2

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
11-3-78	1.00	0-.72	.33	29	10
11-9-78	1.00	0-.52	.21	29	10
11-17-78	1.00	0-.10	.05	21	10
12-11-78	1.00	.04-.17	.07	28	5

Class 5

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-18-78	1.00	0-.21	.13	24	5
10-18-78	1.00	.04-.33	.18	24	5
10-25-78	1.00	0-.17	.10	24	5
10-25-78	.96	.09-.22	.13	24	5

Class 7

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-3-79	1.00	.07-.23	.11	30	10

Class 8

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-3-79	1.00	0-.09	.02	23	10

ADDITION

zzz
+zz

Class 3

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-5-78	1.00	0-.08	.02	12	5
10-12-78	1.00	0-.20	.06	10	5

Class 4

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
09-28-78	1.00	.04-.11	.05	28	10
10-04-78	.96	.04-.20	.09	26	5
10-11-78	.96	.08-.16	.10	26	5
09-27-79	.96	0-.09	.05	24	5

Class 5

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
09-27-79	.95	.11-.28	.23	19	5
09-28-78	1.00	0-.54	.32	24	10
10-04-78	1.00	0-.52	.31	27	5
10-11-78	.71	.18-.59	.32	24	5
11-01-78	1.00	.04-.27	.17	26	5
11-01-78	.92	0-.17	.10	25	5
11-08-78	1.00	.08-.28	.18	25	5
11-15-78	.92	.09-.27	.15	24	5
02-07-79	.96	0	0	28	2
02-21-79	.90	.10-.14	.12	26	2
02-28-79	.88	0-.045	.02	25	2

ADDITION

zzz
+zzz

Class 2

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
12-15-78	1.00	.04-.18	.09	28	5
01-12-79	.96	.08-.23	.12	27	5

Class 4

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
03-08-79	1.00	0-.07	.04	28	3
05-03-79	.96	.04	.04	28	3

Class 5

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
01-10-79	.96	.13-.30	.22	24	5
01-17-79	.96	0-.33	.22	25	5
01-31-79	1.00	.04-.14	.09	28	2
02-07-79	.91	.10-.22	.16	28	2
02-14-79	.95	.125	.125	29	2
02-21-79	.85	0-.23	.12	26	2
03-08-79	1.00	.04-.07	.06	27	3
03-15-79	.96	.04-.16	.11	26	3
03-22-79	1.00	.08	.08	26	3
03-29-79	.96	.04-.25	.10	25	3
04-19-79	1.00	0-.08	.04	26	3
04-26-79	1.00	.04-.08	.05	25	3

Class 7

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-17-79	1.00	.07-.26	.15	27	10
10-31-79	1.00	.04-.36	.15	28	10
11-07-79	.89	0-.25	.14	27	5
11-14-79	.96	0-.25	.12	25	5

ADDITION

zzz
+zzz

(Continued)

Class 8

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-17-79	1.00	0-.25	.15	24	10
10-31-79	1.00	.08-.20	.12	25	5
10-31-79	1.00	.04-.12	.09	25	5

ADDITION

zzzz
+zzz

Class 3

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-12-78	1.00	0-.20	.06	10	5

Class 4

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-04-78	.96	.08-.36	.16	26	5
10-11-78	1.00	.07-.23	.15	26	5
10-25-78	1.00	.04-.21	.12	28	5
11-01-78	1.00	0-.12	.06	26	5
11-08-78	1.00	0-.19	.09	27	5

Class 5

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-04-78	.93	0-.48	.31	27	5
10-11-78	.84	.10-.55	.38	24	5
02-07-79	.84	.11-.19	.15	28	2
02-14-79	.84	.05-.22	.14	29	2
02-28-79	.89	.10	.10	25	2

ADDITION

zzzz
+zzzz

Class 4

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
03-08-79	.97	.04-.15	.10	28	2
05-03-79	.97	.04-.08	.06	28	2

Class 5

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
01-31-79	.97	.08-.30	.18	28	3
03-08-79	.89	0-.17	.08	27	2
03-15-79	.83	.12-.21	.16	26	2
03-22-79	.94	.06-.26	.16	26	2
03-29-79	.94	.11-.19	.15	25	2
04-19-79	1.00	.08-.12	.10	26	2
04-26-79	.93	.05-.14	.09	25	2

Class 7

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
11-28-79	.92	.08-.29	.19	26	5

Class 8

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
11-07-79	.92	0-.13	.08	25	5
11-14-79	1.00	.04-.12	.09	25	5
11-28-79	1.00	0-.19	.08	26	5

ADDITION, THREE ADDENDS

Class 4

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
01-24-79	1.00	.10-.17	.14	29	5
01-31-79	1.00	0-.13	.09	30	5

Class 5

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
01-24-78	.96	.09-.65	.32	24	5
01-31-78	.96	.04-.22	.10	28	5

Class 7

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
11-07-79	1.00	.15-.52	.29	27	10
11-14-79	.96	.08-.46	.23	25	10
11-28-79	.96	.04-.56	.24	26	10

ADDITION, FOUR ADDENDS

Class 3

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-19-78	1.00	0-.33	.165	9	10

Class 4

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-18-78	.97	.18-.46	.32	29	10
10-25-78	1.00	.18-.57	.41	28	10
11-01-78	1.00	.08-.38	.20	26	10
11-08-78	1.00	.11-.41	.22	27	10
01-24-79	1.00	.03-.38	.17	29	10
01-24-79	1.00	.10-.38	.20	29	5
01-31-79	1.00	.03-.23	.15	30	5

Class 5

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
01-24-79	.96	.09-.57	.37	24	10
01-24-79	.89	.25-.63	.47	24	5
01-31-79	.93	.12-.46	.26	28	5

Class 7

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
12-12-79	1.00	.08-.58	.31	24	20
12-20-79	.95	.05-.43	.24	22	20

Class 8

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
11-07-79	1.00	.16-.52	.32	25	10
11-14-79	.96	.08-.44	.28	26	10
11-28-79	.96	.08-.40	.24	26	10

DECIMAL ADDITION

Class 1

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
03-02-79	1.00	.06-.53	.25	32	5
03-09-79	1.00	.03-.20	.11	30	10

Class 5

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
11-29-79	.81	.05-.29	.18	26	5
12-06-78	.95	.06-.17	.09	19	5

Class 6

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
02-21-79	1.00	0-.09	.06	33	5
03-08-79	.94	.03-.10	.07	33	3

SUBTRACTION

zz
-z

Class 2

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
12-01-78	1.00	.05-.42	.21	19	10

SUBTRACTION

zz
-zz

Class 2

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
12-01-78	1.00	0-.53	.27	19	10
12-11-78	.96	.04-.48	.23	28	5

Class 5

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-18-78	.83	0-.35	.18	24	5
10-18-78	.92	0-.32	.20	24	5
10-25-78	1.00	.08-.33	.20	24	5
10-25-78	.92	.09-.41	.21	24	5

Class 7

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-03-79	.93	.07-.89	.55	30	10
10-17-79	1.00	.04-.37	.20	27	10

Class 8

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-03-79	.96	0-.27	.16	23	10

SUBTRACTION

zzz
-zz

Class 3

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-05-78	1.00	0-.33	.16	12	10
10-12-78	1.00	0-.20	.12	10	5

Class 4

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
09-28-78	.86	.17-.58	.43	28	10
10-11-78	.96	.08-.28	.20	26	5
09-27-79	.92	.09-.27	.16	24	5

Class 5

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
09-28-78	.66	.12-.82	.66	26	10
10-11-78	.64	.35-.74	.51	24	5
11-01-78	.92	.12-.38	.30	26	5
11-01-78	.88	.05-.86	.38	25	5
11-08-78	1.00	.12-.88	.30	25	5
11-15-78	.88	.00-.76	.27	24	5
11-29-78	.88	.00-.48	.30	26	5
12-06-78	.95	.05-.28	.18	19	5
02-07-79	.98	.125	.125	28	2
02-14-79	.999	.14-.27	.21	29	2
02-21-79	.98	.10-.22	.16	26	2
02-28-79	.88	.00-.45	.23	25	2
03-29-79	1.00	.04-.16	.10	25	2
09-27-79	.84	.19-.44	.30	19	5

SUBTRACTION

zzz
-zzz

Class 2

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
12-15-78	.86	.08-.42	.32	28	5
01-12-79	.96	.12-.65	.31	27	5

Class 4

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-04-78	.92	.04-.46	.25	26	10
03-08-79	.93	.00-.19	.095	28	2
05-03-79	.95	.10-.25	.175	28	2

Class 5

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-04-78	.78	.10-.67	.51	27	10
01-10-79	1.00	.04-.58	.22	24	5
01-17-79	.96	.08-.38	.20	25	5
01-31-79	1.00	.04-.25	.11	28	3
02-07-79	.75	.00-.62	.31	28	2
02-14-79	.98	.05-.33	.19	29	2
02-28-79	.96	.00-.33	.165	28	2
03-08-79	.85	.00-.39	.20	27	2
03-15-79	.92	.00-.08	.04	26	2
03-22-79	.83	.17-.44	.31	26	2
04-19-79	.99	.15-.23	.19	26	2
04-26-79	.97	.09	.09	25	2

Class 7

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-31-79	.75	.09-.57	.37	28	10
11-07-79	.96	.04-.58	.32	27	5
11-14-79	.84	.20-.57	.38	25	5
12-05-79	1.00	.00-.31	.20	26	20

SUBTRACTION

zzz
-zzz

(Continued)

Class 8

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-17-79	1.00	.04-.21	.11	24	10
10-31-79	.96	.12-.29	.21	25	5
10-31-79	1.00	.08-.24	.14	25	5
12-05-79	1.00	.00-.24	.09	25	20

SUBTRACTION

zzzz
-zzz

Class 3

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-05-78	1.00	0-.08	.03	12	5
10-12-78	.90	0-.44	.18	10	5
10-19-78	.89	0-.25	.075	9	10

Class 4

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-11-78	.92	0-.29	.13	26	5
10-18-78	.93	0-.41	.19	29	10
10-25-78	.93	.11-.35	.23	28	5
11-01-78	.96	.16-.40	.27	26	5
11-08-78	.96	.08-.42	.24	27	5

Class 5

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-11-78	.46	.09-.64	.42	24	5
02-07-78	1.00	.18	.18	28	2
02-21-78	.97	.17-.52	.35	26	2

SUBTRACTION

zzzz
-zzzz

Class 4

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
03-08-79	.96	.19-.54	.36	28	3
05-03-79	.97	.08-.37	.23	28	3

Class 5

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
01-31-79	.86	0-.08	.04	28	2
03-08-79	.96	.19-.54	.36	27	3
03-15-79	.82	.15-.30	.25	26	3
03-22-79	.93	.26-.67	.41	26	3
03-29-79	.93	.14-.40	.22	25	3
04-19-79	1.00	.12-.54	.28	26	3
04-26-79	.97	.09-.54	.25	25	3

Class 7

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
11-28-79	.85	.19-.73	.38	26	5

Class 8

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
11-07-79	.88	.09-.41	.24	25	5
11-14-79	.88	0-.45	.22	25	5
11-28-79	.81	.14-.24	.19	26	5

SUBTRACTION OF DECIMALS

Class 1

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
03-02-79	1.00	.09-.78	.48	32	5

Class 6

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
02-21-79	.97	.06-.38	.19	33	5
03-08-79	.94	0-.19	.12	33	3

MULTIPLICATION

z
xz

Class 2

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
02-16-79	.96	.13-.71	.51	25	10
02-23-79	1.00	0-.28	.09	25	10
03-02-79	1.00	0-.67	.35	24	10

Class 4

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
02-07-79	1.00	0-.17	.04	29	10

Class 5

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
02-21-79	1.00	0-.58	.19	26	10
02-28-79	1.00	0-.40	.16	25	15
03-08-79	1.00	0-.41	.12	27	15

MULTIPLICATION

zz
xz

Class 3

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-26-78	1.00	0-.13	.01	8	10
12-07-78	1.00	0-.17	.08	29	4
12-07-78	1.00	.03-.24	.15	29	4

Class 4

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
01-10-79	1.00	.11-.39	.26	28	10
01-17-79	.96	0-.33	.18	28	5
02-21-79	1.00	0-.16	.06	25	4

Class 5

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
03-15-79	.92	0-.63	.39	26	10
03-22-79	.73	.16-.42	.33	26	10
03-29-79	.88	.18-.59	.37	25	10
04-19-79	1.00	.04-.65	.33	26	10
04-26-79	.96	.04-.46	.34	25	10
09-27-79	.95	0-.78	.41	19	10

Class 8

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
12-12-79	1.00	0-.20	.10	15	20

MULTIPLICATION

$$\frac{zzz}{xz}$$

Class 4

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
01-17-79	.97	.07-.52	.29	28	5
02-21-79	1.00	.04-.08	.05	25	3

Class 8

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
12-20-79	1.00	0-.31	.15	16	20

MULTIPLICATION

zzzz
xz

Class 4

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
02-14-79	.93	.04-.33	.19	29	3
02-21-79	1.00	.16-.24	.21	25	3

MULTIPLICATION

$$\begin{array}{r} zz,zzz \\ \underline{\quad xz} \end{array}$$

Class 4

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
02-14-79	.96	.21-.50	.34	29	3

MULTIPLICATION

zzz,zzz
xz

Class 4

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
02-14-79	.89	.19-.50	.36	29	3

MULTIPLICATION

z,zzz,zzz
xz

Class 4

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
02-14-79	.80	.35-.74	.54	29	3

MULTIPLICATION

zz
xzz

Class 1

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
01-18-79	1.00	.04-.26	.11	27	5

Class 4

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
02-28-79	.87	.20-.36	.28	29	4
03-08-7º	1.00	.07-.21	.13	28	3

Class 6

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
09-27-78	1.00	.04-.45	.20	22	10

MULTIPLICATION

zzz
xzz

Class 1

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
01-18-79	.93	.16-.44	.30	27	5

Class 3

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-12-78	1.00	.06-.33	.18	18	10
12-07-78	.97	.04-.25	.16	29	4

Class 4

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
02-14-79	.79	.18-.68	.39	28	10
02-21-79	.96	.08-.64	.32	26	10
02-28-79	.86	.23-.72	.46	29	3
03-08-79	1.00	.04-.39	.25	28	3
03-15-79	.96	.15-.50	.32	27	10
03-22-79	1.00	.08-.65	.40	26	10
03-29-79	.96	.12-.64	.43	26	10
04-19-79	.97	.07-.48	.31	30	10
04-26-79	.97	.07-.43	.25	29	10
09-27-79	.88	.14-.52	.32	24	10

Class 6

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-04-78	1.00	.13-.54	.35	22	10
10-11-78	1.00	.05-.40	.21	20	10

MULTIPLICATION

zzzz
xzz

Class 4

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
02-28-79	.79	.26-.43	.34	29	3
03-08-79	.94	.20-.58	.36	28	4

MULTIPLICATION

zzz
xzzz

Class 1

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
01-24-79	1.00	.04-.62	.35	26	10

MULTIPLICATION OF DECIMALS

Class 1

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
03-13-79	.08	.62-1.00	.89	32	10
03-21-79	1.00	.03-0.25	.13	32	10

Class 6

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
02-28-79	1.00	.04-.29	.19	29	10
03-08-79	1.00	.09-.27	.17	33	4

DIVISION

$z\overline{)zz}$

Class 3

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-19-78	1.00	0-.22	.10	18	10
11-30-78	.96	.10-.30	.18	21	3

Class 4

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
11-15-78	.96	.16-.44	.36	26	10
11-29-78	1.00	0-.48	.22	23	5
11-29-78	.99	.30-.87	.57	23	5
12-06-78	1.00	.04-.37	.21	27	7
12-06-78	.91	.18-.35	.28	27	3
01-31-79	.93	.07-.75	.31	30	10
02-07-79	1.00	.03-.34	.19	29	10

Class 6

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-25-78	1.00	0-.15	.07	20	10
11-01-78	1.00	.05-.10	.06	21	5

DIVISION

$z\overline{)zzz}$

Class 1

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
02-01-79	.93	0-.23	.15	28	5
03-05-79	1.00	.04-.17	.11	24	3

Class 3

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
11-30-78	.82	.18-.24	.20	21	3

Class 4

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
03-15-79	.56	.41-.93	.67	27	10
03-22-79	.96	.08-.46	.27	25	10
03-29-79	.92	.18-.45	.33	24	10
04-19-79	.93	.22-.48	.32	29	10
04-26-79	1.00	0-.41	.26	29	10

Class 6

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
11-01-78	.95	.05-.40	.20	21	5
11-29-78	1.00	0-.11	.04	19	3

DIVISION

z | zzzz

Class 1

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
02-13-79	1.00	.04-.31	.15	26	5
03-05-79	1.00	.04-.17	.11	24	3

Class 6

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
11-29-78	.95	0-.17	.07	19	3

DIVISION

$zz\overline{)zz}$

Class 3

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
11-30-78	.90	0-.37	.23	21	3

DIVISION

zz⌈zzz

Class 3

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
10-26-78	.94	0-.27	.15	16	10
11-30-78	.48	0-.10	.07	21	3
02-08-79	.92	.13-.38	.22	26	10

Class 6

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
11-08-78	1.00	.09-.32	.25	22	10
11-15-78	1.00	.05-.29	.20	21	10
11-29-78	.95	.11-.17	.13	19	3
01-10-79	1.00	0-.16	.09	18	10

DIVISION

zz⟌zzzz

Class 1

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
02-13-79	1.00	.04-.46	.30	26	5
03-05-79	.93	.24-.51	.37	24	4

Class 3

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
02-15-79	.93	.27-.58	.51	28	5
02-22-79	.79	.08-.31	.20	22	2

Class 6

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
11-29-78	.90	.19-.24	.22	19	3
03-15-79	1.00	0-.46	.26	24	10
03-22-79	1.00	0-.33	.11	30	10

DIVISION

zz⟌zzzzz

Class 3

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
02-15-79	.86	.17-.75	.46	28	2
02-22-79	.82	.11-.39	.30	22	5
02-22-79	.87	.32-.48	.39	22	3

DIVISION

zzz⟌zzzzz

Class 3

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
02-15-79	.71	.44-.55	.49	28	2

DIVISION INVOLVING DECIMALS

Class 1

Date	$\hat{K}$	Range of $\hat{X}_t$	Average $\hat{X}_t$	n	m
03-26-79	.69	.33-.84	.68	28	10
04-23-79	1.00	0-.24	.11	33	10

SECTION IV

ABSTRACTS OF SELECTED RESEARCH STUDIES

Adams, E. N. On scoring a mastery learning system control test. Journal of Computer-Based Instruction, 1974, 1, 50-58.

The state of learning of a student is modeled as M or N, to agree with the outcomes mastery and non-mastery made of a control test. The quality of the test item is characterized by two error parameters, equal to the probabilities of errors of testing, Types I and II. A scoring algorithm is specified based on the probabilistic theory of inference. A bootstrap system for determining error parameters is described, capable of providing continuously improving estimates of error parameters from analysis of individual learner performance data, and it is shown to converge to true values of the parameters in a system for which the underlying model is valid. A Baysean approach is used to determine parameters. In contrast to the classical model of symmetrical random error, in their model positive and negative errors are unrelated. The possible application of the ideas to control test scoring is discussed.

In this paper the author acknowledges Emrick with whom he prepared a report on the implications of a dichotomous states model for the reliability of test scores of conventional and pass-fail varieties in the mastery learning situation.

Aims, D. A Markov model for predicting performance on criterion-referenced tests. Southwest Regional Laboratory Technical Memorandum.

A Markov model for predicting performance on criterion-referenced tests is presented. The model is expressed mathematically as a function of a transition matrix (T), a current state vector (V_c), and a future state vector (V_f). The matrix is defined in terms of conditional probabilities; i.e., the probability of making a transition to a specific future performance state given data pertaining to the student's current performance state. Performance is expressed in terms of mastery, a theoretical construct that is defined in the paper. State vectors indicate either the probability of mastery or the degree of mastery. The current state vector can be computed from available observed criterion test scores.

Three examples are included which indicate how transition matrices may be computed. An example is also provided which shows how the model can be used to predict future performance. Finally, a research application and a management application of the Markov model are mentioned.

Definitions

(V) - State Vector; is a function of probability of mastery or degree of mastery.

(V_c) - Current State Vectors: Exist for all units of instruction that have been completed.

(V_f) - Future State Vectors: can be calculated by using one current state vector and the appropriate transition matrix.

Mastery - is a theoretical construct used to represent the maximum performance level for a specified content objective when performance is measured with a criterion test which makes no assessment errors.

Transition probabilities are based on test scores obtained for a sample of students who have previously completed the necessary units of instruction.

Mathematically, the prediction model is represented by the following matrix equation: $V_{fj} = T_{ij} \cdot V_{ci}$

where V_{fj} is the future state vector for unit j, V_{ci} is the current state vector for unit i, and T_{ij} is the transition matrix from unit i to unit j.

Applications of the Markov Model:

1. To predict the effects of various instructional sequences on subsequent performance.
2. To select an optimal decision strategy, e.g. in business.

A Markov model enables prescriptions to be based on a strategy which maximizes future predicted performance. The method may be evaluated by comparing it with a strategy that maximizes current performance. Evaluation of the two prescription methods would likely involve a direct comparison of actual posttest performance.

Anderson, J., Kearney, G. E., & Everett, A. V. An evaluation of Rasch's structural model for test items. The British Journal of Mathematical & Statistical Psychology, 1968, 21, 231-238.

Rasch's item analysis model for intelligence tests was examined using a 45-item spiral omnibus intelligence test, administered to two samples of 608 and 874 subjects respectively. Five separate hypotheses were tested:

1) Item difficulty indices are independent of the sample on which they are based
2) Indices of item difficulty are not substantially influenced by other items in the test
3) Item difficulty indices are more stable when only items that fit the model are considered
4) Indices of item difficulty are not influenced by the composition of ability groupings
5) Ability level indices are independent of the sample on which they are based.

Findings provide evidence for Rasch's claim that the difficulty level of items and ability level indices are independent of the sample on which they are based.

In the beginning of the paper, a review of Rasch's model is given. It is noted that the model does not lend itself to small samples.

Anderson, T. W. On estimation of parameters in latent structure analysis. Psychometrika, 1954, 19, 1-10.

The latent structure model considered here postulates that a population of individuals can be divided into m classes such that each class is "homogeneous" in the sense that for the individuals in the class the

responses to K dichotomous items or questions are statistically independent. A method is given for deducing the proportions of the population in each latent class and the probabilities of positive responses to each item for individuals in each class from knowledge of the probabilities of positive responses for individuals from the population as a whole. For estimation of the latent parameters on the basis of a sample, it is proposed that the same method of analysis be applied to the observed data. The method has the advantages of avoiding implicitly defined and unobservable quantities, and of using relatively simple computational procedures of conventional matrix algebra, but it has the disadvantages of using only a part of the available information and of using that part asymmetrically.

Baker, F. B. Origins of the item parameters X_{50} and β as a modern item analysis technique. Journal of Educational Measurement, 1965, 2, 167-180.

The purpose of this paper is to bring together the developments relevant to the curve fitting methods of item analysis. The approach is to present the developments in essentially chronological order from its inception in the Binet studies to its modern implementation on digital computers.

The author points out that the modern digital computer has freed us from nearly all constraints due to data processing or computation associated with item analysis, therefore we should not continue to operate under yesterday's limitations. He notes that despite Lawley's paper showing that mental test theory should begin with the specification of the characteristics of the items within an instrument and that subsequent theory should be built upon the item parameters, most of the current mental test theory begins with the test score and ignores the underlying composition of that score. Baker concludes that full advantage of the technological advances can be made only when modern item analysis techniques become an integral part of the total process of test development.

Barcikowski, R. S. The effects of item discrimination on the standard errors of estimate associated with item-examinee sampling procedures. Educational and Psychological Measurement, 1974, 34, 231-237.

A Monte Carlo study was conducted using item-examinee sampling procedures to examine the standard error of estimate for a given test's mean and variance. The main variables considered were test length, item difficulty, and item discrimination. The results indicate that optimal estimates, i.e., smallest standard error, of both mean and variance from a single item-examinee sampling plan may not be possible.

Barcikowski, R. S. A Monte Carlo study of item sampling (versus traditional sampling) for norm construction. Journal of Educational Measurement, 1972, 9, 209-214.

Using a computer-based model of an item trace line, a random sampling experiment concerned with comparing item sample estimates to traditional (examinee) sample estimates of the mean and variance of a distribution of test scores was conducted. The results indicated that the optimal method for estimating a test's parameters may depend on several conditions. As expected, item sampling proved superior to traditional sampling in estimating test means under all conditions. However, with certain test lenghts, ranges of item difficulty, and discrimination, traditional sampling provided better estimates of test variance than did item sampling.

Barcikowski, R. S., & Terranova, C. Item sampling. Unpublished paper.

This paper is concerned with the least number of items which are necessary to provide adequate test norms. This study was designed to consider the item sampling technique under several conditions of item difficulty and item discrimination. The problem was to determine whether the item sampling procedure is better or worse for obtaining estimates of means and variances than the traditional methods of sampling examinees or schools, when the items are of various difficulty and discrimination ranges.

The study was conducted on the IBM 7044 computer. A population of 300 examinees, 50 test items, and a distribution of test scores dependent upon them were considered. To obtain the scores on the computer, an examinee received a score of 1 if he answered an item correctly, and a score of 0 if he answered it incorrectly. The normal ogive model was used and the following assumptions were made:

(1) a unidimensional continuum of the variable of interest
(2) the examinees are distributed normally along this continuum
(3) the items are dichotomous
(4) the items are independent of each other at any given point on the continuum; and
(5) a rectangular distribution of item difficulty and item discrimination.

Several points were made by the authors:

(1) This was not an empirical study. To further substantiate the results, an empirical study would be of interest.
(2) The results indicated that item sampling could be used primarily in the situation where examinees have the same ability level (homogeneous groups) and the discrimination of the items is high. Most norming situations can fulfill these conditions with slight methodological modifications. However, in those classroom situations where there are heterogeneous groups, item sampling procedures may not apply.
(3) A more efficient method of comparison between item sampling and traditional sampling procedures must consider cost of instruments and administration as well as time expended under both procedures, in addition to the variables considered in this study.
(4) This method is not appropriate for speed tests since it assumes that performance of an item is independent of the content in which the item is met.

(5) The use of the method described in this paper should be of use in the systematic study of the optimum number of both examinees and items necessary to provide adequate test norms.

Bejar, I. I. Assessing the unidimensionality of achievement tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

The author presents two procedures which seem useful for detecting violations of the unidimensionality assumption made by latent trait models without requiring factor analysis of inter-item correlation matrices. The procedures require that departures from unidimensionality be hypothesized beforehand by sorting the items within the test into content categories. This is usually possible in achievement tests where several content areas or objectives are included in the test. The rationale of the technique is based on the fact that if the latent space is unidimensional then performance on the test, or a subset from it, should be the same regardless of which items are included in the test. These two procedures are illustrated with data from two biology mid-quarter exams.

Bell, A. I. A comparison of three equating procedures on the certifying examination for primary care physician's assistants. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

A question has been raised concerning the stability of the reference group for the certifying examination for primary care physician's assistants, since the educational background of the students who are entering the training programs is higher today than it was when the test was normed.

To answer the question about the ability of the current examinees, four equating methods were used:

(1) Rasch equating procedure
(2) Linear raw equating procedure
(3) A "short-cut" version of (2), and
(4) Item statistics equating procedure.

The Rasch procedure was best able to answer questions about the items and the examinees.

Benson, J. A comparison of the one- and three-parameter logistic models on measures of test efficiency. A paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, 1979.

The purpose of this study was to empirically compare tests developed by the one- and three-parameter logistic models in terms of relative efficiency. The data employed were a 50 item verbal analogy test from a sample of 5000 high school students. The procedures included item selection, computation of ability estimates based on cross-validation samples, and a relative efficiency comparison of the two tests. The

results indicated that cognitive tests developed using the one- and three-parameter logistic models did not differ in terms of their relative efficiency for most ability groups.

Benson, J. A comparison of three types of item analysis in test development using classical and latent trait methods. Unpublished paper.

This study was designed to empirically compare the precision and efficiency of a cognitive test constructed by three different methods of item analysis. Classical item analysis, factor analysis and the Rasch logistic model were used in the construction of 15 and 30 item tests, replicated for samples of 250, 500, 995 examinees. The study was designed in three phases:

(a) item selection
(b) double cross-validation of the selected items, and
(c) statistical analyses of the test item characteristics

The results of the analyses showed that there were no apparent differences in the types of tests produced by the three methods of item analysis with regard to the precision of measurement.

It was noted that 30% of the items on the 15 item tests and 60% of the items selected on the 30 item tests were common to each item analytic method. Therefore, as test length increased the three methods tended to select the same items.

Thus, the question to consider is: Should practitioners in the field of measurement spend their time learning to use the Rasch model to develop cognitive norm-referenced tests knowing the extra work and sophistication of knowledge required to effectively use the Rasch procedures? With the criterion of internal consistency as a measure of test superiority, it appeared from this study that time spent factorially developing tests, or if computer facilities were not available, the use of classical item analysis procedures seem more than adequate for good test construction.

However, internal consistency which is an integral part of classical test theory (and may be biased since it was derived from the classical model) may not be a fair and sufficient criterion. The relative efficiency formula (Lord) was not derived for any specific test development theory; therefore, relative efficiency estimates should be applicable to any test development technique.

Generally, the results indicated that the Rasch test was superior to the two tests based on classical test theory for students of average to high ability. The two tests based on classical test theory, however, were superior in efficiency to the Rasch developed test for very low and very high ability students, Thus, the test constructor must ask himself, for which segment(s) of the examinee population is the test intended to discriminate?

Benson, J., Crocker, L. M., & Ware, W. B. A comparison of three types of item analysis in test development using classical and latent trait methods. Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, 1978.

This study was designed to empirically compare the precision and efficiency of a cognitive test constructed by three different methods of item analysis. Classical item analysis, factor analysis and the Rasch logistic model were used in the construction of 15 and 30 item subtests; replicated for samples of 250, 500 and 995 examinees. The study was designed in three phases:

(a) item selection
(b) double cross-validation of the selected items, and
(c) statistical analyses of the test and item characteristics.

The results of the analysis showed that there were no apparent differences in the types of tests produced by the three methods of item analysis with regard to the precision of measurement. However, in terms of test efficiency, the results indicated substantive differences in the tests produced by the three methods of item analysis for students with varying ability levels.

Comparisons of relative efficiency for the 30 item tests showed that the tests based on classical test theory were superior to the Rasch developed test for very low and very high scoring examinees, and the Rasch developed test was more efficient for average to high scoring examinees on the verbal aptitude college admissions subtest used in this study.

Berk, R. A. A consumers' guide to criterion-referenced test item statistics. Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, 1978.

The author evaluated sixteen item statistics recommended for use in the development of criterion-referenced tests. Two major criteria were considered:

(1) practicability in terms of ease of computation and interpretation and
(2) meaningfulness in the context of the development process.

Most of the statistics were based on a comparison of performance changes (pretest-posttest) or differences (uninstructed-instructed) between criterion groups. Descriptions and critiques of the difficulty, discrimination and homogeneity indices are presented in the form of a "consumers' guide". It was found that the relatively complex indices offered few if any advantages over the ones that were manually calculable and easily interpretable.

Berk, R. A. A critical review of content domain specification/item generation strategies for criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

The author notes that recently some of the leading proponents of criterion-referenced tests have argued that the objectives-based approach to specifying content domains for purposes of test construction is inadequate. The arguments focus on the subjectivity involved in compositing those specifications. Along with this criticism it has been charged that traditional item construction procedures used to write items from the specifications are also ambiguous.

Six strategies for specifying content domains have been proposed as alternatives to the objectives-based approach of the 1960's and early 1970's. Their effectiveness is assessed in terms of the extent to which they provide an unambiguous domain definition and explicit rules for constructing items such that any two test-makers would produce identical items from the same specifications.

The six strategies that were critically reviewed in this paper are:

(1) amplified objectives
(2) IOX test specifications
(3) item transformations
(4) item forms
(5) algorithms
(6) mapping sentences

The general and technical characteristics of these strategies were surveyed. There was a particular emphasis on the major components, the rule structure by which the content domain is linked to the test items, the type of item domain from which the item sample is generated, and the projects and content domains to which the strategy has been applied. A comparison based on these characteristics is made.

A rating system was also devised to evaluate the strategies according to eight factors of practicability. These factors are:

(1) clarity
(2) simplicity
(3) availability
(4) development time
(5) development cost
(6) adaptability
(7) domain appropriateness
(8) practicability.

Item transformations, item forms, and algorithms offered the most rigorous and precise specifications, while amplified objectives, IOX test specifications, and mapping sentences tended to be the most practical.

Berk, R. A. Item sampling from finite domains of written discourse. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

A sampling methodology is proposed for determining the lengths of tests designed to assess the comprehension or readability of written discourse. It is an extension of Bormuth's work on transformational analysis with a criterion-referenced measurement framework. The sampling units specific to this type of analysis and alternative sampling models are examined. Guidelines are provided for computing sample size and selecting the sample of sentences to which the transformational rules can be applied. A table of sample sizes for a given set of conditions is presented to aid the practitioner in utilizing the sampling theory.

Derived from theoretical linguistics by Bormuth, transformational analysis represents one of the first major attempts to operationally define domains of written discourse such that the generation of finite item domains is possible. Sets of explict rules are employed to transform

sentences selected from textual material into items that measure comprehension of those sentences. The rules provide a direct link between the content and the items.

The sampling units suggest at least three alternative sampling plans:

(1) one-stage cluster sampling of passages
(2) two-state or subsampling of passages and sentences within passages
(3) one-stage simple random sampling of sentences.

The author notes that there are five major factors that affect sample size:

(1) the statistic used as the basis for interpreting the test results
(2) the estimation error associated with the sample statistic
(3) the level of confidence one can place in that statistic
(4) the magnitude of the statistic
(5) the domain size.

Each of these is described in relation to sentence sampling.

It is noted that the most popular procedure to draw a random sample from a domain involves the selection of a set of random numbers equal to the sample size. A simpler and more convenient approach equivalent to this is recommended for sentence sampling It is called systematic sampling.

Despite the utility of the sampling procedures. Berk notes that several technical issues remain unresolved. E.G., how many and what type of transformations should be performed on a sentence that is sampled from a given domain have not been delineated. Other concerns related to the operations of transformational analysis have also been expressed.

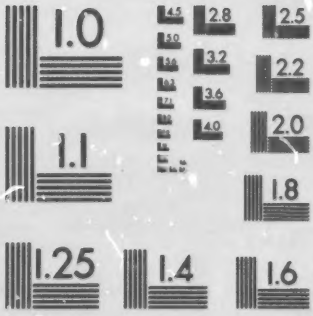
Berk, R. A. Some guidelines for determining the length of objectives-based criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, 1979.

The author examined four factors essential to determining how many items should be constructed or sampled for a set of objectives. These include:

(1) importance and type of decisions to be made with the results
(2) importance and emphases assigned to the objectives
(3) number of objectives
(4) practical constraints.

Within the context of related research on cut-off scores and reliability, it was recommended that between five and 10 items per objective be employed for most classroom decisions and between 10 and 20 items be used for school, system, and state level decisions. Specific guidelines are provided for the use of teachers and evaluator. An illustrative application is included.

Berkson, J. Maximum likelihood and minimum chi-square estimates of the logistic function. Journal of the American Statistical Association, 1955, 50, 130-162.



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS
STANDARD REFERENCE MATERIAL 1010a
(ANSI and ISO TEST CHART No. 2)

Although the minimum chi-square and the maximum likelihood estimates are identical in many situations commonly encountered in statistical practice, there are also some situations (which occur not infrequently in practice) in which they are not identical. For finite samples, the estimates may differ in their distributions, and the question arises, "Which is the better estimate?" The author states that although conjectural opinions favor the maximum likelihood estimate, little or nothing is reliably known which will provide an answer to this question. This article reports the results of one series of experiments Berkson performed in order to clarify this problem.

Bernknopf, S., & Bashaw, W. L. An investigation of criterion-referenced tests under different conditions of sample variability and item homogeneity. Paper presented at the annual meeting of the American Education Research Association, San Francisco, 1976.

This study was designed to examine whether or not traditional procedures concerning item selection and reliability are both applicable and appropriate for criterion-referenced tests. It was also designed to examine traditional procedures and those designed especially for CR testing in relation to test variance and item homogeneity. Specifically, the following questions were formulated:

(1) How are traditional and criterion-referenced item selection techniques interrelated?

(2) How are traditional and criterion-referenced reliability indices interrelated?

(3) How are traditional and criterion-referenced item selection techniques affected by subject variability and test homogeneity?

(4) How are traditional and criterion-referenced reliability indices affected by subject variability and test homogeneity?

The results of the present study indicate that the construction of criterion-referenced tests can be greatly facilitated by item analysis procedures such as phi, and the application of traditional reliability estimates such as KR-20. The procedure of trying out test items on a group of examinees consisting of masters and non-masters is recommended.

Berry, K. J., Martin, T. W. & Olson, K. F. A note on fourfold point correlation. Educational and Psychological Measurement, 1974, 34, 53-56.

The authors present formulas for a modification of Pearson's fourfold point correlation. Possessing always - attainable limits of ± 1 and intermediate values operationally interrpretable in terms of proportionate reduction in error of estimation.

Besel, R. A mastery-learning test model. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972.

The author derived a test model for analyzing criterion-referenced test data. All individuals tested were assumed to be in either the mastery or the non-mastery states. Bayes formula was used to compute

state probabilities. Methods for estimating prior probabilities were described. Two statistics or "decision variables" were computed: Probability of mastery for an individual and proportion in mastery for an instructional group. The relationship between two tests was represented as an adjustment matrix. The interpretation of adjustment matricies in terms of instructional effectiveness and the validation of learning hierarchies was discussed.

Boruch, R. F., & Wolins, L. A procedure for estimation of trait, method and error variance attributable to a measure. Educational and Psychological Measurement, 1970, 30, 547-574.

The procedure described in this paper is based on explicit models for multitrait-multimethod data. The formal models are related to models implicit in the classical Campbell and Fiske (1959) presentation. The authors offer the following discussion of this paper.

Given a number of allegedly different methods of measuring some (other) number of allegedly different attributes or traits of an observational unit, one can use an analytic procedure to examine specific aspects of this situation. The procedure is essentially a quantification and expansion of the systematic assessment of multitrait-multimethod matrices as described by Campbell & Fiske.

Given three or more methods of measuring three or more traits, one can assess the extent to which the observation is influenced by the particular method. The results of the analysis provide a means of determining the extent the methods produce bias. This can be done for each trait. Each method bias may enter more heavily into measurement of one trait than it enters into the measurement of the other.

This can be done for each method of measurement. Specifically, one can establish the degree to which measures of the same traits or attributes are related to one another after one has accounted for method biases. In the Campbell-Fiske nomenclature this is known as discriminant validation.

The statistical procedure requires that one hypothesize a linear model to account for the data. The inferences described in the two items above are, of course, conditional on this model being true. The procedure developed has the distinct advantage of allowing one to assess the goodness of fit of the model. This is done either through the use of a chi-square statistic, or in the sense of desirability of attributes of a solution.

One can also assess the extent to which individual differences contribute to the observations, independent of the particular method-trait combination used in measurement. This is important insofar as one would like to examine global factors such as General Reputation, etc. in assessment of individuals.

The results of this procedure seem somewhat less equivocal than conventional factor analysis procedure since the rotation is uniquely specified by the design. It shares with the conventional procedures problems of nonuniqueness, convergence to local minimum and under determination of factors. To the extent that the various criteria proposed earlier can be used, the solutions appear to be adequate summarizations of the data.

Brennan, R. L. The calculation of reliability from a split-plot factorial design. Educational and Psychological Measurement, 1975, 35, 779-788.

This paper treats the question, "How should one estimate the reliability of schools (or classrooms)?" The author reviews the use of variance components in the estimation of reliability (or generalizability) coefficients in a split-plot factorial design (SPF) with persons nested within schools.

Through the use of variance components from the SPF design, he derives estimates of reliability for schools and for persons within schools. He then compares the reliability for persons within schools from a SPF design with the reliability for persons from a randomized block design. Finally, he compares the reliability for schools from a SPF design with the reliability for school means from a randomized block design.

Brennan says that a randomized block design is a repeated measures design in which the interaction in the population of person i with item j is confounded with experimental error.

Brennan, R. L. Final report: Psychometric methods for criterion-referenced tests. Albany, New York: The Research Foundation of the State University of New York, 1974.

The first four chapters of this report primarily provide an extensive, critical review of the literature with regard to selected aspects of the criterion-referenced and mastery testing fields. Major topics treated include:

(1) definitions, distinctions, and backgrounds
(2) the relevance of classical test theory
(3) validity and procedures for test construction
(4) test reliability.

Chapter V provides a treatment of criterion-referenced and mastery item analysis and revision procedures when items are scored in the classical correct/wrong manner.

Brief summaries to each chapter are provided.

Brennan, R. L. Some applications of generalizability theory to the dependability of domain-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Using the basic principle of generalizability theory, a psychometric model for domain-referenced interpretations is proposed, discussed and illustrated. The author points out that this approach is applicable to numerous data collection designs, including the traditional persons-crossed-with-items design, which is treated extensively here.

It is shown that the appropriate error variance for domain-referenced interpretations is what Cronbach and others call $\sigma^2(\Delta)$, rather than $\sigma^2(\delta)$, which is the error variance for norm-referenced interpretations. Also, it is shown that two indices can be developed that reflect

the dependability of a domain-referenced testing procedure. These indices are interpretable in several complementary ways, and they are easily compared with generalizability coefficients.

In this paper, consideration is given to theoretical results as well as to estimation procedures, illustrative examples, extensions to multiple facet designs, and recommendations for researchers and practitioners.

Brennan, R. L., & Lockwood R. E. A comparison of two cutting score procedures using generalizability theory. ACT Technical Bulletin No. 33, April, 1979. Also, a paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, 1979.

The authors note that Nedelsky and Angoff have suggested procedures for establishing a cutting score based on raters' judgments about the likely performance of minimally competent examinees on each item in a test. In this paper, generalizability theory is used to characterize and quantify expected variance in cutting scores resulting from each procedure. Data for a 126-item test are used to illustrate this approach and to compare the two procedures. Finally, consideration is given to the impact of rater disagreement on some issues of measurement reliability or dependability. Results suggest that the differences between the Nedelsky and Angoff procedures may be of greater consequence than their apparent similarities. In particular, the restricted nature of the Nedelsky (inferred) probability scale may constitute a basis for rejecting this procedure in certain contexts. It is important to note, however, that the numerical results reported in this paper are for a single study, only. As such, the authors caution, they do not form a sufficient basis for passing judgment on either the Nedelsky or the Angoff procedure.

Brownless, V. T., & Keats, J. A. A retest method of studying partial knowledge and other factors influencing item response. Psychometrika, 1958, 23, 67-73.

A method of studying the problem of correction for guessing and other problems associated with behavior in the test situation is described and an illustrative example presented.

They assume the following:

(1) At the first administration, all responses are either known correctly, guessed, or "known" incorrectly.

(2) At the second administration, all responses are either known correctly, guessed, "known" incorrectly, or repeated from memory.

(3) No person who knows the correct answer at the first administration will guess at the second.

(4) No person will learn an incorrect response between administrations.

Capper, J. Approaches to standard setting in competency based education: Framework for a composite model. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

The author describes several approaches to setting performance standards and then provides a framework for a composite model of standard-setting in minimum competency testing programs.

Existing standard-setting models include:

(1) Referencing standards to the performance of others on the same measure. Both Glass & Millman have described an approach to standard-setting which is referenced to the performance of a population of examinees (adults).

(2) Referencing standards to the performance of others on a related measure.

(3) Referencing standards to teacher judgments. Teachers who are familiar with a group of students identify those students whom they are 'certain' are either masters or non-masters. Scores for the two groups are graphed and cut-off point is set where the two curves intersect. Jaeger describes a similar approach which involves comparing the performance histories of examinees with their performance on the measure for which the standard is being set. Zieky & Livingston describe an approach wherein teachers render judgements regarding students they consider borderline in a particular subject area.

(4) Glass describes an "Operations research" approach to setting standards. An example of this would be to find out what levels of performance on a criterion-referenced test (eg. reading) optimize performance on an external criterion (eg. success in college or job). This approach is closely allied to Jaeger's notion of inferring performance on a sample of domain tasks to an ultimate criterion.

(5) Focus on items. Nedelsky, Ebel, & Angoff devised variations on this basic approach of focusing their attention solely on the items.

(6) Approaches based on analysis of classification errors. In this instance, a standard has previously been set, but is reviewed in light of misclassifications resulting from this standard. (Jaeger, Hambleton & Novick, Emrick, Kriewall, Shepard and Glass.) The goal is to minimize the errors of classification.

Related issues:

(1) Should the standard be set with or without the use of actual performance data? Wiersma & Jurs, and Zieky & Livingston, favor establishing the criterion prior to any actual measurement. Glass and Popham support the use of performance data in the setting of standards.

(2) Should normative performance data be used? Shepard recommends this but suggests that when making decisions for newly developed measures, the normative information can be derived from similar existing measures. Glass opposes this.

(3) Should absolute or comparative values be used for setting standards? Glass supports the notion of using comparative data. Shepard concedes that there may be some instances where absolute standards are required (eg. H.S. grad. requirements) and where these standards can be clearly determined.

FRAMEWORK FOR A COMPOSITE MODEL OF STANDARD-SETTING

I. Description of desirable criterion-referenced test characteristics to be considered by persons responsible for selecting or developing a measure.
    a. relevance of content to the ultimate task.
    b. homogeneity of items.

c. item congruence with content description.
d. representativeness of item sample.

II. Factors that should be considered in setting performance standards:
1. Analysis of test content:
a. difficulty of the content:
b. importance of mastery of the content for acquiring subsequent skills.
2. Empirical evidence:
a. summary of criterion groups' performance on the measure.
b. comparison of test results with teacher judgments.
c. Comparison of criterion groups' performance with other relevent groups performance on criterion measure.
d. information regarding students' abilities to transfer their skill to related areas.
e. comparison of students' rate of retention to their performance on the criterion measure.
3. Analysis of classification errors:
4. Decision consequences:
These may be generated based on field test performance data.
5. Preferences of various groups - eg. parents, students, teachers, subject matter experts, employers and college & university faculty.

Cohen, J. Weighted chi square: An extension of the kappa method. Educational and Psychological Measurement, 1972, 32, 61-74.

This article presents a very general method for the study of m-way tables of proportions or frequencies (where m is one or more) in which the investigator's a priori hypotheses about the cells are expressed numerically and used as weights. These weights are then used in k_w, an index of hypothesized association, and also in a test of its significance, weighted chi-square, which thus utilizes as relevant information the investigator's hypotheses.

The system Cohen describes is a direct outgrowth of work which was initiated to provide a coefficient of agreement for nominal scales. The measure which was proposed, kappa, is simply the proportion of agreement for the N cases placed in the k categories by the two judges, corrected for chance agreement. Standard error formulae for significant testing and setting confidence limits were also presented, and since for large samples k is approximately normally distributed, statistical tests and estimates take the familiar classical form.

Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 1968, 70, 213-220.

A previously described coefficient of agreement for nominal scales, kappa, treats all disagreements equally. A generalization to weighted kappa (k_w) is presented. The k_w provides for the incorporation of ratio-scaled degrees of disagreement (or agreement) to each of the cells of the k x k table of joint nominal scale assignments such that disagreements of varying gravity (or agreements of varying degree) are

weighted accordingly. Although providing for partial credit, k_w is fully chance corrected. Its sampling characteristics and procedures for hypothesis testing and setting confidence limits are given. Under certain conditions, k_w equals product-moment r. Although developed originally as a measure of reliability, the use of unequal weights for symmetrical calls makes k_w suitable as a measure of validity.

Cook, L. L. & Hambleton, R. K. A comparative study of item selection methods utilizing latent trait theoretic models and concepts. Paper presented at the annual meeting of NCME, San Francisco, 1979.

Using the three-parameter logistic test model and the concept of score information curves, the purposes of this investigation were:

(1) Provide some background on information curves for items and tests;

(2) Using a typical item pool compare the score information curves for five item selection methods:

- (a) random
- (b) standard
- (c) middle difficulty
- (d) up and down
- (e) maximum information

(3) Compare the merits of several item selection methods for producing a scholarship exam and a test to optimally separate examinees into three ability categories.

In all cases, the item selection methods based on either the random selection of items or the use of classical item statistics produced results inferior to those produced by methods utilizing latent trait model item parameters. The appropriateness of each method was situation specific. If maximum information is required at only one point on an ability continuum, a method which chooses items that maximize information at this particular point will be the best. If information is required over a wider range of abilities, methods involving averaging the information values across the ability levels of interest or choosing items in some systematic method that considers each point of interest on the ability continuum appear to be promising.

The authors point out that although only a limited number of methods and testing situations have been investigated, the results indicate that it may be possible to prespecify item selection methods that are situation specific and will enable a practitioner to develop a test quickly and efficiently without going through a lengthy trial and error process.

A variable not considered in this study was the effect of the item pool on the successful application of the methods investigated. The authors suggest that it is possible that different results might have been found for item pools containing items with differing characteristics. The authors caution that further research which consider other types of information based item selection methods as well as method-item pool interaction is certainly necessary before a complete set of generalizable guidelines can be developed.

Cooley, W. W. Explanatory observational studies. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

The task of this paper was to describe the state of the art in the design and analysis of research studies involving relationships among variables. The focus is on observational studies. These are studies that inquire into the learning and development of human beings in the natural environments in which these processes occur. These studies are multivariate and longitudinal, recording the variances and covariances of events and individual differences as they occur and unfold. There is particular focus on observational studies that are intended to be explanatory rather than descriptive.

Cooley describes the following major requirements for sound observational studies. If these requirements are met, he believes they will yield consistent, convincing, useful explanations of educational phenomena.

(1) the sampling framework, which affects the generalizability of of the observed relationships

(2) the theoretical mode, which describes the hypothesized causal structure of the variables under consideration,

(3) the statistical procedure, which is used to analyze the network of observed relationships for the purpose of establishing the plausibility of the theoretical model and estimating its parameters.

The author concludes that more convincing causal models are needed. The challenge, he says, is considerable, but the efforts will be more productive than continuing to conduct meaningless quasi-experiments, or averaging all of the t-tests they may have produced. Cooley states that as educational research is guided by increasingly valid models of educational phenomena, and the causal networks that are currently operating are better understood, we may eventually draw close to that "complete covariate" that Crobnach and his associates have shown is necessary for interpretable quasi-experiments.

Cooley, W. W. & Leinhardt, G. Design and educational findings of the instructional dimensions study. Paper presented at the annual convention of the American Educational Research Association, Toronto, 1978.

In 1974 the National Institute of Education (NIE) launched a major survey of the impact of specific educational practices on student development. It was called the Instructional Dimensions Study (IDS). This report briefly outlines the design and presents some of the results of the IDS. Specifically, it focuses on identifying effective classroom processes in regular classroom settings.

The results of the study indicate that what goes on in a classroom has definite impact on what students learn during the course of an academic year. At least one-fourth of the variation in achievement gain is due to differences in classroom processes. What is less clear is precisely which classroom practices make what kind of a difference with whom.

The major generalization the authors make from the analyses is that the most useful construct in explaining achievement gain is the opportunity that the children had to learn the skills assessed in the achievement test, especially as represented by the measures of overlap between the curriculum and the posttest.

The results of IDS show that students are much more likely to answer correctly if they have been directly taught the material covered by a test and if they have been exposed to the test format. The finding emphasizes the importance of being able to justify the form and content of tests used in evaluations.

Another implication of the importance of the opportunity construct is that program evaluations that do not include information on how time is allocated or on the degree of overlap between curricula and test, run the risk of attributing instructional effectiveness to specific programs or ways of teaching when it is really a matter of differences in opportunity.

What the data also seem to indicate is that there are many different ways of teaching and that no one way--individualized or grouped--is superior. No one technique of instruction is clearly associated with disastrous outcomes or successful ones.

This is consistent with that fact that there was no clear evidence regarding the superiority of individualized instruction for compensatory education.

The preliminary results of this study support the idea that the emphasis in instruction should be on the cognitive rather than on the managerial.

Cox, R. C. Item selection techniques and evaluation of instructional objectives. Journal of Educational Measurement, 1965, 2, 181-185.

The major conclusions of this study are:

1) Statistical selection of items from the total item pool has a biasing effect on the selected tests. The proportion of items in the selected tests which measure certain instructional objectives is unlike the proportion of items in the total item pool which measures the same objectives. The selected tests are not representative of the total item pool in this respect.

2) Statistical selection of items from the total item pool operates differentially for male and female groups. When the statistical data obtained from the female tryout group is used to select tests from the total item pool, the results differ from those obtained using the male tryout group. The structure of the selected tests, as indicated by the taxonomical structure of the items, differs from the male and female groups.

Crawford, C. R. Item difficulty as related to the complexity of intellectual processes. Journal of Educational Measurement, 1968, 5, 103-107.

Intellectual processes defined in both Bloom's (1954) taxonomy and by the Committee on Student Appraisal (1962) are considered to be

hierarchical. Because of this hierarchical principle, it has been argued that items measuring the more complex processes are, by their very nature, more difficult than items measuring the less complex processes.

The purpose of this study was to investigate the relationship between item difficulty and complexity of intellectual processes presumably measured by multiple-choice items when knowledge is not held constant.

The results indicate that the order of difficulty level was, in every analysis except one, statistically different from the order of complexity. This suggests that there is not necessarily a direct relationship between the complexity of intellectual processes and the difficulty of items which purportedly measure them. This finding is consistent with Guttman's [In Lazarsfeld (Ed.), Mathematical thinking in the social sciences, Glencoe, Ill,: Free Press, 1954, p. 283] statement:

> "There is some danger of confusing the notion of degree of complexity with that of difficulty. If we say that subtraction is more complex than addition, we do not mean by this that subtraction is necessarily more difficult than addition. Complexity and difficulty have no necessary connection with each other in our theory."

Crehan, K. D. Item analysis for teacher-made mastery tests. Journal of Educational Measurement, 1974, 11, 255-262.

The focus of this study is on item selection for teacher-made mastery tests. The author questions whether teacher-made tests resulting from various item selection techniques differ when evaluated by appropriate methods of estimating criterion-referenced reliability and validity.

Six item techniques are compared.

Eighteen volunteer junior and senior high teachers wrote behavioral objectives and parallel items for each of the original items. The entire pool of items was administered to two classes before and after instruction and to two other classes only after instruction.

Pairs of tests developed by each of the six methods were derived. Estimates of test reliability and validity were obtained using responses independent of the test construction sample.

No specific selection method resulted in consistently higher reliability rankings; but the modified Brennan and Cox-Vargas methods consistently resulted in higher observed validity rankings.

The author notes that generalizations of this study are limited because of the nonrandom observations. However, the author assumes that criteria empolyed for reliability and validity are appropriate for evaluation of teacher-made mastery tests. Crehan questioned whether the magnitude of improvement in test validity of objective item selection over teacher selection is worth the necessary effort on the part of the teacher.

Cureton, E. E. Note on ϕ/ϕ_{max}. Psychometrika, 1959, 24, 89-91.

Cureton gives formulas for a descriptive statistic related to the fourfold-point correlation but having always-attainable limits of $\pm$ 1.

Cureton, E. E. Reliability of multiple-choice tests is the proportion of variance which is true variance. Educational and Psychological Measurement, 1971, 31, 827-829.

Frary (Educational and Psychological Measurement, 1969, 29, 359-365) presented an analysis which seemed to show that classical weak true-score theory does not apply to multiple-choice tests. Cureton showed that the difficulty with Frary's derivation is that the guessing score is not separated into a true component and an error component.

Cureton, E. E. The stability coefficient. Educational and Psychological Measurement, 1971, 31, 45-55.

The author noted that the formula he previously presented (Educational and Psychological Measurement, 1958, 18, 715-738 and Educational and Psychological Measurement, 1965, 25, 327-346) for the stability coefficient was essentially the same formula given by Remmers and Whistler (Journal of Educational Psychology, 1938, 29, 81-92). Although the formula is correct, both his derivation and the one given by Remmers and Whistler were slightly defective. A derivation which the author believes to be more nearly correct was presented in this paper together with some further discussion.

Divigi, D. R. A new index for the accuracy of a criterion-referenced test. Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, 1978.

One aim of criterion-referenced testing is to classify an examinee without reference to a norm group. Therefore any statements about the dependability of such classification ought to be group-independent also. A population-independent index is proposed in terms of the probability of incorrect classification near the cut-off true score. The compound binomial model leads to the conclusion that a criterion-referenced test is more reliable if the item difficulties are unequal.

Dowing, S. M. & Mehrens, W. A. Six single-administration reliability coefficients for criterion-referenced tests: A comparative study. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

The purpose of this study was to compare several criterion-referenced reliability coefficients to the Kuder-Richardson estimates and to each other. KR 20 & 21, the Livingston, the Subkoviak and two Huynh coefficients (K, k) were computed for a random sample of 33 criterion-referenced tests. The Subkoviak coefficient yielded the highest mean

value; Huynh's Kappa yielded the lowest. The Huynh K and $\hat{k}$ coefficients were highly positively correlated with the Kuder-Richardson 20 and 21 coefficients, and with each other; the Livingston and the Subkoviak indexes were highly correlated with each other. A two-factor principle components solution suggested that only the Subkoviak coefficient measured a test characteristic that differed from the classical (KR) internal-consistency coefficients.

The data for this study were 33 achievement exams which represent a random sample of objective format (3 to 5 option multiple-choice) criterion-referenced (mastery) exams from undergraduate teacher education classes, medical school classes, and state-wide assessment tests. The number of exam items ranged from 5 to 143, with a mean of 38.9 items. The number of subjects taking these tests ranged from 5 to 1110 with a mean of 209.9. Each of the six reliability coefficients was computed for each exam.

Results support the usefulness of KR-21 with criterion-referenced examinations. The authors also believe that their result suggests that the Livingston Coefficient may be more useful for criterion-referenced reliability than its critics have allowed.

Duncan, G. T. An empirical Bayes approach to scoring multiple-choice tests in the misinformation model. Journal of the American Statistical Association, 1974, 69, 50-57.

This article develops multiple-choice test scoring rules, concentrating on Bayes rules and their frequency theory analogs, empirical Bayes rules. Conditions are given for empirical Bayes estimates to lie in the probability simplex. The misinformation model is considered in detail. It is shown that ranking by raw scores is equivalent to ranking by Bayes scores when the loss function increases with error and the sampling distribution has the monotone likelihood ratio property. Application of the techniques is made to data from a multiple-choice test given to students of an elementary statistics course.

The misinformation model postulates that a given examinee has knowledge of the correct response to p (p=0,1,...n) items and perceives a wrong answer as correct on w (w=0,1,...n-p) more items. The examinee then has misinformation about w items.

Ebel, R. L. The case for non-referenced measurements. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

In building his case for norm-referenced measurements, Ebel divides this paper into four sections. The first lists differences between norm and criterion-referenced tests. These include differences in the age and development of the two test forms, in the kind of information they provide, in their sampling of tasks, in the educational purposes they serve, in the range of achievement levels they measure, in the range of schools for which they are appropriate, and in the conceptions of learning they imply.

Similarities of the test types are pointed out in the second section. Some of these include the fact that the items used in the two tests are indistinguishable. The kind of tasks to be included in each can be specified precisely. The territory and the boundaries of the domain of achievements from which particular tasks are to be selected can be defined with all the precision that is necessary for either type of test. Ebel contends that both kinds of tests yield scores that differ from pupil to pupil, although a test might be built to show no score variance. The author emphasizes his belief that score variance is not irrelevant to any test of achievement. Another point of similarity between norm and criterion-referenced tests has to do with norms. Ebel notes that norms are involved in establishing the criteria of achievement on which criterion-referenced tests are based.

The third section of this paper is Ebel's case against easy items. The author states that these items should not be thrown out provided two requirements are met:

1) the item unquestionably tests an achievement of unquestionable importance

2) the number of items answered correctly is more important than its relative value, that is its precentile rank or stanine or z-score.

The fourth section lists some of the advantages Ebel sees of norm-referenced tests.

1) They assess the pupil's broad general level of knowledge and understanding of a subject, not his mastery of a few particulars.

2) they reflect common nation-wide goals for learning, not unique local goals.

3) They assess achievements at all levels of excellence and mediocrity. They do not focus primarily on minimum essentials.

4) Because each item can test a different aspect of achievement, they provide a broader and more representative sampling of achievements.

5) They are consistent with the view that achievement in learning is a matter of more or less, not of everything (mastery) or nothing.

6) They provide a single score that concisely summarizes a pupil's general level of achievement, not an extended inventory of things learned or not learned.

7) They are primarily useful for summative, not formative evaluation. They indicate how successful the pupil's efforts to learn have been; how successful the teachers efforts to foster that learning have been.

8) They imply that the primary responsibility for successful learning rests with the pupil, not with the instructional delivery system.

Ebel, R. L. The relation of item discrimination to test reliability. Journal of Educational Measurement, 1967, 3, 125-128.

In this paper, Ebel maintains that in order to achieve high reliability in a test with a given number of items, one must write or select items that are high in discrimination as measured by D, the upper-level index of discrimination. Data are presented to support this position.

Eignor, D. R. Statistical suggestions for the study of cognitive structures. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

The primary purpose of this paper was to offer some viable statistical techniques to be used in the study of cognitive structures. The procedures developed by Hubert and his associates (Schultz and Baker) offer viable techniques for testing the structures generated out of tasks utilized in the study of cognitive structures. The procedure for scaling the proximity data to have metric properties appears manageable when viewed theoretically, but practical concerns still exist.

The author notes that one further area of research should be considered. The research concern has to do with the situation when the researcher wants to measure one individual's cognitive structure and then compare this to a content structure matrix or an expert's cognitive proximity matrix. This is an issue, Eginor states, that must be dealt with if the measurement procedures for cognitive structure are ever going to replace objective tests as measures of higher order objectives. There are two levels of concern:

1) Which of the tasks (word association, tree construction, F-sort, similarity judgment) should be used with one individual, and

2) Which of the scaling procedures, S. C. Johnson's hierarchical clustering, Kruskal's or Ramsey's multidimensional scaling and Waern's graphing technique, if any, is amenable to individual solutions? As an instance of the first level of concern, it would appear that while it is reasonable to utilize a word association task to generate proximity data on an individual level, use of the F-sort appears unreasonable. A matrix of zeros and ones for an individual is generated in the latter case and the matrix is likely to yield nonmeaningful results.

In summary, Eignor states that research needs to be done on which tasks, and subsequently which scaling or graphing procedures, are best suited for measuring an individual's cognitive structure.

Eignor, D. R. & Hambleton, R. K. Effects of test length and advancement score on several criterion-referenced test reliability and validity indices. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, 1979.

The purpose of this study was:

1) to report the relationships between test lengths and several reliability and validity indices for a fixed cut-off score (80%) in five domain score distributions, and

2) to report the relationships between advancement scores and several reliability and validity indices for several test lengths in five domain score distributions.

Five figures show the relationships between test length and decision consistency, kappa, decision accuracy, predictive validity, and efficienty, respectively, for each of the five domain score distributions under consideration. A number of observations and/or cautions concerning the use of the figures are offered; and suggestions for further research and development are offered.

Epstein, K. I. & Knerr, C. S. Criterion-referenced test interpretations of "classical" measurement theory. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

The purpose of this paper is to remind the practitioner that more than statistics and measurement theory are required in order to interpret test results meaningfully, and to provide two examples which illustrate the importance of considering the entire testing situation in making inferences about a particular test.

This paper suggests that many of the results obtained when "classical" techniques are applied to criterion-referenced tests, particularly in the context of mastery learning, are perfectly reasonable, interpretable, and should be expected.

Estes, G. D., Colvin, L. W. & Goodwin, C. A criterion-referenced basic skills assessment program in a large city school system. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

The Phoenix Union High School System has developed a basic skills assessment program in the areas of Reading and Mathematics. Procedures and considerations used in developing the system are discussed. Preliminary results on validity and reliability of the assessment instruments are presented. An example of criterion-referenced test development using traditional item analysis, reliability, and validity procedures is provided.

Goals and objectives are stated for both Reading and Mathematics.

Procedures used to develop their instruments are outlined. They were very similar to those normally used to develop norm-referenced tests including item difficulty and discrimination values which were used to identify the test items to retain, revise or delete from the item pool.

KR-21 reliability estimates were calculated for their reading and math tests.

Concurrent validity estimates were calculated for their reading and math tests.

A summary of student performance on their math and reading tests is presented.

They claim that students have steadily progressed toward mastery of all areas such that 94.8% of the class of 1972-76 has mastered all the reading areas, and 80.2% of the class of 1973-77 has mastered all of the math areas with almost a year and one-half left before graduation.

Everitt, B. S. Moments of the statistics kappa and weighted kappa. The British Journal of Mathematical and Statistical Psychology, 1968, 21, 97-103.

This paper considers the mean and variance of the two statistics, kappa and weighted kappa, which are useful in measuring agreement between two raters, in the situation where they independently allocate a sample of subjects to a prearranged set of categories.

$$\text{kappa} = \frac{\text{proportion of observed agreements - proportion of expected agreements}}{1 - \text{proportion of expected agreements}}$$

Kappa may be interpreted as the proportion of agreement over and above that expected by chance. It can be shown to have almost the same value as the product-moment correlation coefficient for the dichotomous case. The advantage of kappa appears to be that it is more intuitively reasonable, and also that it leads to weighted kappa, which takes into account the relative seriousness of the different types of disagreement which can arise between the two observers. The mean and variances of kappa and weighted kappa are considered.

Faggen, J. Decision reliability and classification validity for decision oriented criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

In the first section of this paper, a statistical model to study the relationships between reliabilities and validities of content-bound criterion-referenced test item pools is developed. Two item parameters are defined; a_i is the probability of a master answering item i correctly and b_i is the probability of a non-master answering item i incorrectly. Two psychometric properties of the item pools--decision reliability (R_D) and classification validity (V_C)--are developed. The decision-reliability of an item pool is defined as the probability that a randomly selected examinee who is administered two randomly constructed n-item tests is classified as either a master on both tests or as a non-master on both tests. The classification validity measures the degree to which masters score at or above the cut score on a given test and non-masters score below the cut score.

In the second section of this paper, a guide is presented to enable practitioners--curriculum developers, test constructors, and evaluators--to generate the decision reliability matrix and the corresponding classification validity matrix associated with a heterogeneous item pool. The use of Bayesian methods is explored to provide estimates of the several parameters of interest. In addition, the entire theory and procedures are applied to a mathematics item pool which is currently monitoring student progress in a self-pacing format. The predictions made in the first section are borne out successfully. In particular, the data clearly supported the useful relationship that R_D provides an upper bound to V_C.

In the third and final section, several additional problems which need to be pursued with respect to domain-referenced item pool testing are described.

Forsyth, R. A. & Spratt, K. F. Measuring problem solving ability in mathematics with multiple choice items: the effect of item format on selected item and test characteristics. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, 1979.

The major purpose of this study was to investigate the effects of the modified item format on selected item and test characteristics. Specifically, the difficulty and discrimination of such items and the reliability of a test composed of such items were examined relative to the more common format. A limited evaluation of the validity of a test composed completely of such items was also undertaken. In addition, since multiple choice items measuring mathematics problem solving frequently use the "Not given" or "None of these" alternative, and since the research related to the effect of such alternatives on items and test characteristics is not conclusive, the effects of this type of alternative in conjunction with the two item formats discussed were also investigated.

The results of this study suggest that the set-up format yields more difficult and less discriminating items than the traditional format. These generalizations are limited by the particular examinee and item populations that were studied. It is possible, for example, that both item formats may perform equally well with high school students.

The authors note that an extremely limited evaluation of the validity of items using the set-up format was undertaken in this study. A more extensive evaluation would have required information related to the test-taking strategies and problem solving strategies that students use with this type of item. Such data would also be useful for evaluating the validity of the more traditional items. Future studies in this research area should attempt to gather information related to these strategies.

Frase, L. T. The demise of generality in measurement and research methodology. Paper presented at the Center for the Study of Evaluation, Winter Invitational Conference on Measurement and Methodology, Los Angeles, 1978.

In this paper the author does three things. First, he expresses some optimism about the decline of superficial analyses in testing and research methodology. However, he states that there is a need for caution lest the tendency toward precision makes one lose sight of the broadly adaptive characteristics of human behavior that cut across learning tasks. Second, he reviews a model of learning skills that might be used to link specific performances to test items, and to link test items to instructional methods. He notes that this is only one model that could provide a domain focus for item writing, but it has a strong research base and it communicates with the kinds of activities that are encouraged in reading comprehension programs. Finally, he reviews some complexities of inferring processing activities from performance on test items. This analysis leads him to believe that a great deal can be learned about tests and about people by concentrating on the strategies that test-takers use to arrive at the information (or data base) that supports test-taking performance. He notes that test-taking performance is often governed by factors (like motivation and world knowledge) which have little to do with the nominal task presented by a test item.

Frase concludes that practitioners sharpen their conception of learning theory, that they analyze the skills entailed in specific subject matters, and that they explore ways to relate these skills to instructional methods and outcomes. Data he reviewed suggest that not only the idealized conceptions of what cognitive processing is be explored, but also the adaptive behaviors of test-takers that simplify and convert nominal stimulus materials into something other than they are intended to be.

He suggests that it might be true that part of the reason for test score declines is that the background knowledge of students who now take standardized tests has changed, and that some cultural backgrounds do not support concern for how one scores on a test. If so, Frase concludes, then it is all the more important to characterize and explore partial and optional representational processes, since these have to do with motivational and cultural factors which are logically prior to many of the performances that our test items attempt to measure.

Gagne, R. M. Observing the effects of learning. Educational Psychologist, 1975, 11, 144-157.

The author states that the effects of learning are typically observed in test situations which need to be evaluated in terms of both construct and validity. The model empolyed by information-processing theories of learning and memory is proposed as a source of construct validity which provides a rationale for assessing the effects of several different phases of the learning-memory process. As for content validity, the suggestion is made that criterion-referenced measurement be achieved by precise analysis, description, and representation of the criterion through job-sample" testing, which would avoid the apparent narrowness of coverage of "domain-referenced" item forms.

Gardner, P. L. Test length and the standard error of measurement. Journal of Educational Measurement, 1970, 7, 271-273.

The author shows that under very general conditions, the standard error of measurement estimated from the Kuder-Richardson formula 20 and Kuder-Richardson formula 21 leads to Lord's observations that the standard error of measurement of a test is directly proportional to the square root of the number of items on the test.

Grosse, M. E. An application of the Rasch model to common person equating. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

This paper describes the use of the Rasch model to equate pairs of examinations taken by the same group of examinees. The examinations had no items in common, so this is an example of common person equating.

When traditional psychometric methods were applied to an examination, neither scores nor reference groups suitable for the purpose of recertification could be obtained. The Rasch model provided a two step

solution to these problems. First, it provided a rationale for combining measures derived from different sections of the examination. Second, it provided a method for discovering and adjusting the differential difficulty of the subspecialty examinations. The final result was a single common ability measure for each candidate. A reference group of satisfactory size could then be identified for the purpose of establishing a norm referenced standard for acceptable performance on the examination.

Grosse, M. E., Wright, B. D., Shumacher, C. F. An application of the Rasch model to common person equating. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

This paper describes the use of the Rasch model to equate pairs of examinations taken by the same group of examinees. The examinations had no items in common, so this is an example of common person equating. The analyses were conducted for the recertification examination of the American Board of Plastic Surgery.

The first issue addressed with the Rasch model was the question of the unidimensionality versus multidimensionality of the underlying trait measured by the 5 examination books. The second issue addressed with the Rasch model was the intrinsic, unknown difficulty of the subspecialty examinations.

In summary, when traditional psychometric methods were applied to an examination, neither scores nor reference groups suitable for the purpose of recertification could be obtained. The Rasch model provided a 2-step solution to these problems. First, it provided a rationale for combining measures derived from different sections of the examination. Second, it provided a method for discovering and adjusting the differential difficulty of the subspecialty examinations. The final result was a single common ability measure for each candidate. A reference group of satisfactory size could then be identified for the purpose of establishing a norm referenced standard for acceptable performance on the examination.

Gumbel, E. J. Bivariate logistic distributions. Journal of the American Statistical Association, 1961, 56, 335-349.

The author notes that the logistic distribution closely resembles the normal one. Both are symmetrical. In this paper two logistic bivariate distributions are studied. In both cases the curves of equal probability density are not ellipses, the regression curves are not linear and the conditional expectations are limited. The first distribution analyzed with the help of the bivariate moment generating function is asymmetrical and therefore departs considerably from the normal one. The coefficient of correlation is constant and equal to one half. The second bivariate logistic distribution is symmetrical. The regression curves are linear in probability scale and the coefficient of correlation varies in the interval $\pm$ .30396.

Gustafsson, J. E. Testing and obtaining fit of data to the Rasch model. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

In this paper the author treats problems and procedures in assessing and obtaining fit of data to the Rasch model. The assumptions embodied in the model are made explicit and it is concluded that statistical tests are needed which are sensitive to deviations such that more than one item parameter would be needed for each item, and such that more than one person parameter would be needed for each person. Statistical goodness-of-fit tests, based on conditional maximum likelihood estimates of the item parameters, which can detect these two kinds of deviation are presented. Common sources of deviation are also identified, as are the tests needed to detect them. Problems in the use of statistical tests to assess fit are discussed and some investigations of power are presented. In relation to a distinction between use of the Rasch model as a criterion and as an instrument, the treatment of the goodness-of-fit problem in different measurement contexts is discussed. Finally it is concluded that items which can be identified as misfitting should not be routinely excluded to obtain fit to the model; instead other actions should often be taken such as grouping of the items into homogeneous subsets.

Haladyna, T. M. An investigation of full- and subscale reliabilities of criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1974.

The author notes that classical test theory has been rejected for application to criterion-referenced (CR) tests by most psychmetricians due to an expected lack of variance in scores and other difficulties. The present study was conceived to resolve the variance problem and explore the possibility that classical test theory is both appropriate and desirable for some types of CR tests. Both a rationale and empirical evidence were offered to support the practice of using unrestricted samples to estimate full- and subscale reliabilities of CR tests using classical procedures. However, reservations were expressed concerning the reliability of these subscales.

Haladyna, T. & Roid, G. The stability of Rasch item difficulty and student achievement estimates for a criterion-referenced test. A paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, 1979.

The purpose of this study was to determine the applicability of the Rasch model to criterion-referenced (CR) testing. Since a CR test is sensitive to instruction, item difficulties tend to vary as a function of instruction. In this study, CR tests data were subjected to a Rasch item analysis under three sample conditions:

1) uninstructed students only
2) instructed students only
3) combined samples.

The authors make the following conclusions.

1) Difficulty is not stably estimated for these CR tests. The Rasch item statistics tend to fluctuate as a function of instruction.

2) The mean-square-fit index appears to be poorly estimated regardless of sample condition. Thus, these results suggest that MSF not be considered useful information in CR analysis.

3) Examinee achievement estimates tend to remain stable regardless of which sample is used to establish these estimates. These results attest to the robustness of the Rasch model for providing estimates of student achievement based on a set of difficulty estimates, despite the fact that these estimates vary from sample to sample.

Hambleton, R. K. A review of testing and decision-making procedures for selected individualized instructional programs. ACT Technical Bulletin No. 15, August 1973.

The first purpose of this investigation was to provide a description of the testing models that are currently being used in selected individualized instructional programs. Three programs were studied:

1) Individually prescribed instruction
2) program for learning in accordance with needs,
3) mastery learning.

The author provides an introduction for each instructional model. This includes a brief history of the program, the content areas covered, and an indication of the extent of implementation. Also, a description of each instructional paradigm and details on the testing model are provided. An attempt is made to pinpoint the decision points in each model, spelling out the consequences of the various possible actions in relation to each of the possible true states of nature.

A second purpose of this paper was to compare the three programs and the four component parts of the testing model; namely, selection of a program of study, criterion-referenced testing on the unit objectives, assignment of instructional modes, and final year-end assessment.

A final purpose was to briefly outline several promising lines of research in connection with the testing methods and decision procedures for individualized instructional programs.

Hambleton, R. K. Testing and decision-making procedures for selected individualized instructional programs. Review of Educational Research, 1974, 44, 371-400.

The author notes that the successful implementation of an individualized instructional program depends, in part, upon the availability of appropriate testing and decision-making procedures to monitor student progress. In this paper, Hambleton has attempted to describe and to compare the testing models of three of the best known and widely adopted instructional programs: IPI, Project PLAN, and Mastery Learning. In addition, on the basis of a review of the models, he has outlined several important lines of research that could contribute significantly to the quality of testing and decision-making within the context of these and other individualized instructional programs.

Hambleton, R. K. & Cook, L. L. Some results on the robustness of latent trait models. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

The purpose of this paper was to study the goodness-of-fit" of the one-, two-, and three-parameter logistic models. The authors studied, using computer-simulated test data, the effects of four variables: Variation in item discrimination parameters, the average value of the pseudo-change level parameters, test length, and the shape of the ability distribution. Artificial or simulated data representing departures of varying degrees from the assumptions of the three-parameter logistic test model were generated and the "goodness-of fit" of the three test models to the data was studied.

The results of the computer simulations were:

Level of variation in discrimination parameters:

1) For the values studied in the paper, using discrimination parameters as item weights contributed very little to the proper ranking of examinees.

Level of pseudo-chance level parameters:

2) With the 20-item tests, the three-parameter model was considerably more effective at ranking examinees correctly in the lower half of the ability distribution. Correlations were about .08 higher ($\sim$.75 to $\sim$.83) in the uniform distribution of ability and about .08 higher in the normal distribution ($\sim$.65 to $\sim$.73). The improvement in the average absolute difference in rank order was about 13.

3) With the forty-item tests, the three-parameter model was also somewhat more effective at ranking examinees correctly in the lower half of the ability distribution. Correlations were about .04 higher in both ability distributions. The improvement in the average absolute difference in rank order was about 8. The reduction in effectiveness of the three-parameter model weights was to be expected with the longer tests. Gulliksen noted the insignificance of scoring weights when the test gets longer and test items are positively correlated.

4) For examinees in the upper half of the ability distribution, and for the data sets studied, the number rights score was about as effective as the more complicated scoring weights used in the two-and three-parameter models.

Shape of the Ability Distribution:

5) As expected, correlations tended to be higher for the uniformly distributed ability scores.

Test Length:

6) Increases in correlations due to doubling the length of the test were observed. Again, as expected they tended to be rather small.

Hambleton, R. K. & Eignor, D. Guidelines for evaluating criterion-referenced tests and test manuals. Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, 1978.

The authors note that the scope and number of criterion-referenced tests available to potential users is impressive, but that the quality

of these tests varies tremendously and so it is very important for potential users to review available tests carefully before making their selections.

The primary purpose of this paper was to propose a set of guidelines for evaluating criterion-referenced tests and test manuals. The guidelines should be useful to both users and developers of criterion-referenced tests. A secondary purpose was to report on their use of the guidelines with eleven commercially available criterion-referenced test batteries.

The guidelines represent the authors' own biases about what is important technical information for users to have in making informed decisions about the quality of criterion-referenced tests.

Questions were generated by the authors asking themselves "What questions would we want to answer before making a decision to use a criterion-referenced test in a particular situation?" Questions were organized around ten broad categories: Objectives, Test Items, Administration, Test Layout, Reliability, Cut-off scores, Validity, Norms, Reporting of Test Score Information and Test Score Interpretations. Several questions for each category are listed.

Eleven criterion-referenced tests were selected for review. The primary purpose of this article was to ascertain the extent to which these tests met the authors guidelines. The evaluation of each test is reported as well as how the tests as a group meet each of the guidelines.

The authors note that their guidelines are offered only to serve as a "catalyst" for further discussion and debate. Their use of the proposed guidelines to evaluate eleven criterion-referenced tests was intended to

1) demonstrate that the proposed guidelines were workable,
2) highlight areas where considerably more or different work on the part of test developers is needed.

Individuals with suggestions for improving the guidelines were encouraged to write the authors.

Hambleton, R. K. & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

The authors synthesize some of the thinking in the area of criterion-referenced (CR) testing (current in 1973) as well as provide the beginning of an integration theory and method for such testing. They view criterion-referenced testing from a decision-theoretic point of view; thus approaches to reliability and validity estimation consistent with this philosophy are suggested. In order to improve the decision-making accuracy of CR tests, a Bayesian procedure for estimating true mastery scores is proposed. This Bayesian procedure utilized information about other members of a student's group (collateral information), but the resulting estimation is considered to be criterion-referenced rather than norm-referenced since the student is compared to a standard rather than to other students. The authors contend that in theory, the Baye-

sian procedure increases the "effective length" of the test by improving the reliability, the validity, and the decision-making accuracy of the criterion-referenced test scores.

Harms, R. A. The development, validation and application of an external criterion measure of achievement test item bias. Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, 1978.

This paper outlines the need for basis, development, validation and application of an external criterion measure of test item bias.

The criterion measure developed is based on the sole and generally accepted presupposition that fair test items are test items that are just. From this initial presupposition, a criterion measure of item bias based on an operationalization of Rawls' theory of justice, justice as fairness, is developed, validated and applied by student raters to items selected from the 1972 TASK II A. Finally, the by-item results and implications of this application are included and discussed.

Principle I - Within a defined usage (contract) test items should not "turn-off" students so that they are unable to do as well as their abilities would indicate.

Criteria for Principle I:

A) An item must not contain any information that could be offensive to a student's religion or culture.

B) An item must be non-sexual in the sense that it must not be designed to offend either sexual group.

C) An item must not include depictions of any group that are degrading or stereotypic.

D) An item must not portray groups as unequal in ability.

E) An item must not, by content or form, cause students to be "turned-off" so that they may not do as well as they are able.

Principle II - Within the domain of a defined usage (contract) for any given test item, all students should have equality of opportunity to respond.

Criteria for Principle II

F) The content of the test item must not reflect information and/or skills that may not be expected to be within the educational background of all students or groups that take the item.

G) The content of the item must not contain any clues or information that could be seen to work to the benefit of any group or either sex.

H) The content of the item must be shared to all and not specific to any one group or sex.

I) Group specific language, vocabulary or reference pronouns (he, she, etc.) must not be included in items.

Implications - Analyzing test items for bias is most useful in the test construction phase when the test is easily amenable to change. The included criterion measure has been systematically applied in several such applications with encouraging results. It appears to the author, however, that the measure is very possibly conservative and its application will lead to the selection of a larger number of items as biased

than post-hoc statistical procedures. However, as it is easier to rewrite abberant items than to speculate on their effect on scores, such conservatism is likely to be both long-term efficient and beneficial.

The author expects to continue this research by applying five of the published analytical item bias estimation procedures to the TASK items previously examined by the criterion measure. The results of this comparison will be used in a comparative concurrent validation of these analytical estimators.

The author advocates an ongoing test development and item writing-tool that may be used as a benchmark measure of test item bias. [He's not advocating an additional item bias estimation procedure or a replacement one].

Hayek, L. C. An empirical study of properties of mastery models. Smithsonian Institution, April, 1979.

This study presents and empirically evaluates, a probabilistic model for mastery assessment. Characteristics of the estimation procedure are examined and bias of estimates, asymptotic results and goodness of fit problems are discussed. The overall findings suggest that not only is total sample size a major concern for the user, but also average frequency per response vector and number of items were found to be potential problem areas.

Haynes, J. L. & Walker, C. Establishing criteria for objective mastery--art or science? Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

The purpose of this paper is to examine some methods of establishing criteria for mastery for objective-referenced tests.

Recognizing the importance of valid criteria, the Florida Department of Education conducted a series of studies to determine:

1) the validity of the established criteria for masters,

2) whether an empirical basis for establishing criteria for mastery could be identified.

The studies were conducted in conjunction with Florida's statewide assessment program, which administers a state-developed, objective-referenced test in reading, writing and mathematics to students in grades three, six and nine. The three studies are described briefly in this paper.

The first study involved <u>teacher judgment of mastery</u>. If teachers identify students as masters or non-masters and the same classification occurs based on which students meet the established criteria, a high degree of concurrent validity would be established. This presupposes that teacher judgment about student mastery is itself valid and reliable. The findings indicate that any criteria will result in a high percentage of misclassifications when compared to a global teacher judgment about student mastery or non-mastery of a subject area.

The second study involved <u>teacher judgment of item achievement</u>. This study asked teachers to apply their general knowledge of students' competencies to estimate the difficulty of selected items which were part of the assessment instrument. As a general approach to setting

objective criteria, this method seemed to add little information for the amount of work involved. Since teachers would probably continue to use a relatively small range, most mastery scores would be n-1. If that will be the outcome, it would be much easier to set n-1 as the criteria without expending teacher effort to confirm it.

The third study, called one-above study, involved students which were identified by their teachers as being superior students in either communication skills or mathematics. Each student completed the communication skills or math section of the regular assessment booklet for the grade immediately below. This process of administering the test to superior students in the grade above seems to hold the most promise, both as an approach to establishing criteria and as a means of identifying weak items.

Helmstadter, G. C. A comparison of Bayesian and traditional indexes of test item effectiveness. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 1974.

The purpose of this study was to apply Bayes Theorem to item analysis of test questions and to compare the resulting indexes of item effectiveness with the traditional index of test item discrimination when "high" and "low" knowledge groups are defined in three different ways. Empirical data in this study were the responses to items on a pre and post test from 43 students in a multivariate statistics course and from 55 students in an adolescent psychology course.

Bayes theorem was applied and led to the development of three separate indexes of item effectiveness as follows:

1) the probability that a subject knows the content material given the correct response was selected;

2) the probability that he does not know the material given that the incorrect response was selected,

3) the probability that a correct decision will be made about the examinee's knowledge of the content given the results of performance on that item.

These three item characteristics plus a classical item discrimination index, were then computed for each item. Three variants of these four item characteristics were obtained by defining "high knowledge" and "low knowledge" groups in different ways.

Intercorrelations among the twelve different indexes derived for each item were computed over the items within each of the two separate tests. The medians of the intercorrelations among ways of defining groups and the medians of the intercorrelations among the different item indexes were calculated. A principle axis factor analysis with varimax rotation was applied to each intercorrelation matrix, thus obtaining an independent factor structure for each of the two classes.

The data suggest that the common practice of defining "high" and "low" knowledge groups in terms of scores on the posttest only is questionable. They further indicate that there may be two quite distinct types of effective achievement test items: those that indicate that the examinee knows the material and those that indicate that the examinee does not know the material. Thus, another common practice - that of

using a single index of item discrimination - is also called into question.

Hill, R. K. Use of the Rasch model to solve data problems encountered by the California assessment program. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, 1979.

This paper presents four problems that the staff of the California Assessment Program, the statewide testing program for the State of California, found difficult to solve. Each of these problems proved to be readily solvable when techniques being developed by advocates of the Rasch model were applied. The purpose of this paper was not to present new approaches for using the Rasch model, but to demonstrate that the approaches that have been developed already have great practical significance and should be disseminated and used more widely by practitioners.

Hills, J. R. Using empirical data to set cutoff scores. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, 1979.

The author notes that the 1974 edition of Standards for Educational and Psychological Tests takes a clear stand on the setting of cutoff scores. But that presentation was written a decade ago. The author reviews some of the developments since then. This includes the work of Cronback & Snow in which they point out that investigators should be looking for at least interactions of the fourth order; the work of Millman who starts with the true score; the work of Hambleton & Novick who suggest that one consider not only whether an error is made in classifying a student, but which kind of error is more serious; the work of Fhaner modified by Wilcox who suggest that the cutting score be set in terms of an indifference zone. In Wilcox's development we must set the desired probability of making a correct decision; the work of Davis & Diamond which used Bayes's Theorem to calculate the lowest acceptable cutting score. Finally, the author discusses the work of Visco who started from the observed score and chose not to use Bayes's theorem. Visco started with the observed score and a specified confidence interval and solved for the lowest true score that would be in the interval. Visco also turns the problem around and considers setting cutting scores for nonmastery.

The author notes that none of the examined methods offers support to the notion that in criterion-referenced testing, we need only a few items to decide whether a student has mastered an objective. Hills is concerned that there is not a system for classroom teachers to set soundly the cutting scores on their own tests, or evaluate the soundness of the cutting scores given them by publishers.

Hively, W., Patterson, H. L. & Page, S. A. A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-290.

This paper shows that two quite different approaches to achievement testing converge. One is the strong form of educational behaviorism exemplified by B. F. Skinner's work in the area of programmed instruction. The other is the positivistic approach to mental test theory exemplified by Cronbach's work on "Generalizability." Hively et al maintain that if behaviorists can analyze non-trivial subject matter into well-defined behavioral classes, generalizability theory promises appropriate and powerful measurement models. Osburn (1968) has named this area "universe-defined-achievement testing."

Data are presented from one of the first applications. The subject matter is mathematics.

A general form, together with a list of generation rules, precisely define the set of all test items which may be taken to represent the diagnostic category. The rules for generating such a set of test items is called an "item form." A collection of item forms constitute a "universe" from which tests may be drawn. A "family" of random-parallel tests (Cronbach, 1963) is defined by a sampling plan over a universe of item forms. A "generalizability study" of the test families was conducted. Three tests were generated from each family. The results show that the relative magnitudes of the components of variance display an extraordinarily consistent pattern across the different test families. It is emphasized that these results were obtained on the basis of purely formal content analysis of the subject matter, without any statistical item selection procedures whatsoever.

In addition to estimates of "relative" stability of test scores expressed by intercorrelation coefficients, a measure of the individual's performance with respect to the universe, which does not require comparison to other individuals for its interpretation, is also obtainable. To get this, we may use the within-person variance to estimate a confidence interval for any individual's "true" or "universe" score, given his observed score on a randomly chosen test. The underlying assumption in all of the above is that the between- and within-person variances are independent of one another.

The authors continue to say that given information about how a person responded to a particular test item, we would expect to be able to predict how he would respond to another, randomly-chosen item from the same item form, but not necessarily how he would respond to an item from a different item form. Predictions from one item form to another should depend on how the items forms are related.

Hively concludes that the data lead one to place only moderate faith in the item forms as categories which represent distinct, homogenous classes of behavior and which thus provide the foundation for detailed diagnosis and remediation. By contrast, it seems paradoxical that the total test scores should have been as reliable as they were.

Horn, J. L. Integration of concepts of reliability and standard error measurement. Educational and Psychological Measurement, 1971, 31, 57-74.

The purpose of this paper is to explicate some of the problems implied by the assumptions underlying derivations of various indices or

error of measurement and such coefficients of reliability as the Kuder-Richardson formula 20 and the Kuder-Richardson formula 21, and to indicate some of the practical implications of various proposed solutions.

Horn looks at standard error of measurement and reliability coefficients as they are defined in terms of two random response models. The conclusion is that generally the Kuder-Richardson formula 20 should yield a larger estimate of reliability than the Kuder-Richardson formula 21, and although the difference may be small in many practical situations, the fact of the difference between the two should be kept in mind when considering the standard error of measurement formulae.

In the second section of this paper. Horn looks at some standard error of measurement models. It is noted that different kinds of variability can be represented as "error" in any one of the formulae for reliability or standard error of measurement.

Hughes, F. P. The Rasch model applied to the equating of several examination forms. Paper presented at the joint session of the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, San Francisco, 1979.

The equating of six tests was discussed using data obtained from common-item links and analyzed using the Rasch model. An iterative procedure developed by Drs. Wright and Mead for estimating average test difficulty on a common scale, and for estimating the expected value of shifts for common-item and non-existent links, was described and illustrated using these data. Also, a procedure is suggested for evaluating the quality or coherence of the equating data, and its application to the data in this study is discussed.

Two indices were derived to help identify triads and links that lacked coherence. The author recommends that the distributions of these indices should be investigated so that firm guidelines for identifying triads and links that lack coherence can be established.

Hummel-Rossi, B., & Oldak, R. An empirical comparison of methods for determining reliability of criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

Using a two-part aptitude test administered to 226 students criterion-referenced ability estimates were obtained at selected cut-off scores for Swaminathan, Hambleton, and Algina's kappa coefficient, Huynh's kappa coefficient, Huynh's approximation for kappa, Harris's efficiency coefficient, and Subkoviak's coefficient of agreement. Results indicated similarity between the Swaminathan et al. and Huynh kappa estimates. Huynh's approximation for kappa appeared stable. Subkoviak's coefficient gave high estimates close to the observed consistency of classification. Harris's estimates proved unrelated to the others. The authors note that the size of the reliability coefficient is bound strongly to the method of calculation.

Before one computes the reliability of a criterion-referenced test he must decide on the kind of reliability interpretation needed as the various methods provide different types of information as well as different values. In practice one rarely meets all the assumptions of the various models for estimating criterion-referenced reliability; as one deviates from the models the estimates may become less accurate. Investigation of these coefficients under violations of their model assumptions is suggested.

Hutten, L. R. A comparison of the fit of empirical data to two latent trait models. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

The primary question addressed in this study was how well do empirical data fit the one and three-parameter logistic latent trait models. The author notes that little research has addressed the question of comparable model fit and lists three important questions.

1) Should the practitioner select the Rasch model with one type of data, and the Birnbaum (3-parameter) model for other kinds of data?

2) Is there improvement in model-data fit found by using the three-parameter model, rather than the Rasch model?

3) How can practitioners determine which test model (the one or three-parameter model) best suit their data?

The results of this study indicate that for data having items equal in discrimination, the Rasch model provides better fit to empirical data than the three-parameter logistic model. A practical method for determining equality of item discrimination, using classical point-biserials, was suggested. It was also noted that the maximum likelihood estimate of the discrimination parameter may be inadequate at this time. As improvements are made in the three-parameter estimation methods, a more sensitive estimate of this parameter may be found.

Although the data used in this study were multiple choice in nature, violation of the "no guessing" assumption of the Rasch model did not appear to effect fit of the one-parameter model to data. The maximum likelihood procedure tended to overestimate guessing for this data. This caused reduced model-data fit of the three-parameter model especially in the lower ability range. Generally, guessing was unestimable for this data. The author states that better estimates are needed for both item discrimination and guessing if the three-parameter model is to be used effectively.

Using a factor analytic criterion, the data used in this study were all found to have one general factor which, in all cases, accounted for more than 20% of the test variance. There also appeared to be some improvement of fit to both models for data that showed extremely strong first factor variance.

Although the ability estimates from short tests were good, item estimates from small samples of persons tended not to be so good. This result was especially apparent in estimating item discrimination from small samples.

When the logist program was used with known item parameters, the cost of estimation in one and three-parameter cases was equivalent. In

estimating item parameters simulteneously with ability, the savings found by using the one-parameter model were considerable.

In summary, using costs and fit to test score distributions as criteria, the Rasch model was clearly superior in fit to empirical data than the three-parameter logistic model. It is important to note that other criteria for fit might have been selected which would have shown better fit for the three-parameter model.

The results also show that in the case when item discriminations are quite dissimilar, the three-parameter model demonstrated superior fit to the Rasch model. Research is needed to determine how unequal item discrimination needs to be for the three-parameter model to become more effective.

Finally, it is important to note that the conclusions drawn in this paper are tentative. The project is in midstream; less than half of the projected data sets have been analyzed to date.

Huynh, H. Bayesian and empirical Bayes approaches to setting test mastery scores. Paper presented in an AERA/NCME joint syposium on "psychometric approaches to domain-referenced testing", at the annual meeting of the American Educational Research Association, San Francisco, 1979.

The Bayesian mastery scores as proposed by Swaminathan et al. and the empirical Bayes mastery scores derived from Huynh's decision-theoretic framework are compared on the basis of approximate beta-binomial and real CTBS test data. It is found that the two sets of mastery scores are identical or almost identical as long as the test score distribution is reasonably symmetric or when the true criterion level is high. Large discrepancies tend to occur when this level is low, especially when the test scores concentrate at some extreme scores or are fairly bumpy. However, in terms of mastery/nonmastery decisions, the Huynh procedure provides the same classifications as the Bayesian method in practically all situations. Moreover, the former may be used for tests of arbitrary length and has been generalized to more complex testing situations.

Huynh, H. Budgetary consideration in setting mastery scores. Research Memorandum 79-3, publication series in mastery testing, Educational Research Program, Columbia, South Carolina: College of Education, University of South Carolina, 1979.

A general model along with four illustrations are presented for the consideration of budgetary constraints in the setting of cutoff scores in instructional programs involving remedial actions regarding poor test performers. Budgetary constraints normally put an upper limit on any choice of cutoff score. Given relevent information, this limit may be determined. Alternatively, ways to assess the budgetary consequences associated with a given cutoff score are provided. Such information would be useful in any final decision regarding the cutoff score.

Huynh, H. A class of mastery scores based on the bivariate normal model. Research Memorandum 79-4, publication series in mastery testing, Educational Research Program, Columbia, South Carolina, 1979.

This study touches some aspects of the determination of mastery scores on the basis of the bivariate normal test model. The loss ratio associated with classification errors is assumed to be constant, and the referral success function ranges in the normal ogive family. Alternatively the model also provides a fairly simple way to assess the loss consequences associated with each mastery score. Such information is deemed useful to the test user who may wish to examine these consequences before making a final choice of cutoff score. It is also noted that the model provides a latent trait analysis for testing/measurement situations involving instructed and noninstructed groups, or pretest and posttest data.

Huynh, H. Computation and inference for two reliability indices in mastery testing based on the Beta-binomial model. Paper presented at the annual Southeastern Invitational Conference on Measurement in Education, Greensboro, North Carolina, 1978.

In mastery testing the raw agreement index and the kappa index may be secured via one test administration when the test scores follow beta-binomial distributions. This paper reports tables and a computer program which facilitate the computation of those indices and of their standard errors of estimate. Illustrations are provided in the form of confidence intervals, hypothesis testing, and minimum sample sizes in reliability studies for mastery tests.

Huynh, H. A nonrandomized minimax solution for mastery scores in the binomial error model. Research Memorandum 78-2, publication series in mastery testing, Educational Research Program, Columbia, South Carolina: College of Education, University of South Carolina, 1978.

A nonrandomized minimax solution is presented for mastery scores in the binomial error model. The computation does not require prior knowledge regarding an individual examinee or group test data for a population of examinees. The optimum mastery score minimizes the maximum risk which would be incurred by misclassifications. A closed-form solution is provided for the case of constant losses, and tables are presented for a variety of situations including linear and quadratic losses. A scheme which allows for correction for guessing is also described.

Huynh, H. & Mandeville, G. K. An approximation to the true ability distribution in the binomial error model and applications. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

Assuming that the density p of the true ability θ in the binomial test score model is continuous in the closed interval (0,1), a Bernstein

polynomial can be used to uniformly approximate p. Then via quadratic programming techniques, least-square estimates may be obtained for the coefficients defining the polynomial. The approximation, in turn will yield estimates for any indices based on the univariate and/or bivariate density function associated with the binomial test score model. Numerical illustrations are provided for the projection of decision reliability and proportion of success in mastery testing.

The authors note that a linear constraint is built into this algorithm and thus the accuracy of the fitting may be somewhat disturbed. Another version of the approximation with no such linear constraint will soon be investigated by the authors.

Huynh, H. & Perney, J. Determination of mastery scores when instructional units are linearly related. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

A procedure is described for the determination of mastery scores when instructional units are sequenced in a linear hierarchy. The scheme is based on the principle of backward induction. Using an overall evaluation to estimate the criterion level, the mastery score for the last unit is determined first. On the basis of mastery classification for the last unit, a mastery score is computed for the preceeding unit. The process is continued until a mastery score is obtained for the first unit.

A model relating the performance on the referral task to the criterion level and mastery score is based on the beta-binomial density function. The implication is that all test items are of equal difficulty. (Students with lower test scores are more likely to fail the referral task than those with higher test scores.)

The steps for the computation scheme, which is similar to the technique of backward induction, is presented.

Huynh, H. & Saunders, J. C. III. Accuracy of two procedures for estimating reliability of mastery tests. Paper presented at the annual conference of the Eastern Educational Research Association, Kiawah Island, South Carolina, 1979.

The beta-binomial estimates for the raw agreement index p and the kappa index in mastery testing are compared with those based on repeated testings in terms of bias and sampling stability. Across a variety of test score distributions, test lengths, and mastery scores, the beta-binomial estimates tend to underestimate the corresponding population values. The percent of bias, however, is negligible (about 2.5%) for p and moderate (about 10%) for kappa. Both beta-binomial estimates are almost twice as stable as those based on repeated testings. Though the beta-binomial estimates presume equality of item difficulty, the data presented indicate that even gross departures from equality do not affect the performance of the estimates.

Hymel, G. M. & Gaines, W. G. Emrick's model as a basis for determining an optimal criterion score and a ratio of regret in a mastery testing situation. Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, 1978.

The authors note that Emrick's model is an algorithm which generates a test cut rule or mastery criterion based upon both the properties of a given test and a cost-benefit analysis of possible decision errors. The objectives of this paper are:

(1) to present the basic assumptions and formulation of Emrick's ('71) evaluation model for mastery testing.

(2) to demonstrate the application of the model in determining the most appropriate mastery criterion that should be established for a field-tested achievement instrument.

(3) to suggest an ex post facto application of the model to those situations in which the determination of a mastery criterion and the use of a given instrument have already occurred.

In this paper the authors fail to mention the Wilcox-Harris paper, "On Emrick's 'An Evaluation Model for Mastery Testing'" in the Journal of Educational Measurement, 1977, 3. In the Wilcox-Harris paper it is shown that Emrick's expression for the correlation between a mastery state and item response is correct when exactly half of the examinees are masters or when there is no measurement error, but that in general, his expression is incorrect. It is further shown that Emrick's estimation of the probability of misclassifying examinees cannot be justified when the generally-correct correlation is used.

Ironson, G. H. A comparative analysis of several methods of assessing item bias. Paper presented at the annual convention of the American Educational Research Association, Toronto, 1978.

Test data from two diverse culture groups (1,691 blacks; (1,794 whites) were analyzed to determine the agreement among four methods of bias and two indices intended as validity criteria. The four methods of detecting item bias were:

1) transformed difficulty,
2) discrimination differences,
3) chi-square,
4) item characteristic curve.

The first validity measure was based upon matching groups on ability and the second was based upon classifying each item as traditional (verbal & mathematics) or nontraditional (letter groups, mosaic comparisons, picture number). The test battery was given as part of the National Longitudinal Study of High School Class of 1972 and was composed of 155 items from the six substests listed above.

There appeared to be some bias in the battery as measured by

1) those indices having associated significance tests,
2) a comparison of two white groups showing considerably less bias than the black and white comparison groups of interest.

For the 150 items analyzed, three of the methods (which maintained magnitude but not direction of bias),

1) transformed difficulty,
2) chi-square,
3) ICC, approaches were moderately correlated (1&2 = .239, 1&3 = .370, 2&3 = .485).

There was little agreement between the discrimination differences approach and the others. Upon close examination the percent agreement among the items identified as most biased by each method was moderately low.

Of the two validity indices, the traditional verses non-traditional classification of items had higher correlations with the bias methods; the obtained correlations were (.244; .100; .280; and .414) with the four bias methods respectively. The second validity measure examined difficulty differences for a random sample of 100 blacks matched with 100 whites separately on each subtest score. This index correlated significantly with both the transformed difficulty (.279) and ICC (.316) approaches (with both direction and magnitude of bias preserved). The chi-square index was not included in this analysis because the direction of bias is difficult to maintain.

Finally, the correlations were not high enough to offer conclusive advice to the user on which bias methods to employ. Thus, the author suggests that additional studies need to be done to further evaluate the external validity and comparability of the methods.

Kaiser, H. F. & Michael, W. B. Domain validity and generalizability. Educational and Psychological Measurement, 1975, 35, 31-35.

An alternative derivation of Tryon's basic formula for the coefficient of domain validity or the coefficient of generalizability developed by Cronbach, Rajaratnum and Gleser is provided. This derivation, which is also the generalized Kuder-Richardson coefficient, requires a relatively minimal number of assumtions compared with that in previously proposed approaches.

Kane, M. T. & Brennan, R. L. Agreement coefficients as indices of dependability for domain-referenced tests. ACT Technical Bulletin No. 28, Iowa City: The Research and Development Division, 1977.

The authors point out that a large number of seemingly diverse coefficients have been proposed as indices of dependability, or reliability, for domain-referenced and/or mastery tests. In this paper, it is shown that most of these indices are special cases of two generalized indices of agreement, one that is corrected for chance, and one that is not. The special cases of these two indices are determined by assumptions about the nature of the agreement function or, equivalently, the nature of the loss function for the testing procedure. For example, indices discussed by Huynh, Subkoviak, Swaminathan, Hambleton and Algina employ a threshold agreement, or loss function; whereas, indices discussed by Brennan and Kane and Livingston employ a squared error loss function. Since all of these indices are discussed within a single general framework, the differences among them in their assumptions, properties and uses can be exhibited clearly. For purposes of comparison,

norm-referenced generalizability coefficients are also developed and discussed within this general framework.

Katzenmeyer, C. G., Stewart, D. M. & Quilling, M. R. A model for the development and evaluation of placement tests for objective based curriculum management systems. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1974.

This paper outlines a model for the development and evaluation of a placement test for the Word Attack area of the Wisconsin Design for Reading Skill Development. A thirty-item placement test was constructed and tried out in two elementary schools prior to program implementation. Development strategies and effectiveness of the placement test in minimizing leveling errors are discussed.

A student was considered inappropriately leveled if he mastered 0 or only 1 scale at a level (test down) or mastered all or all but 1 test at a level (test up). With regard to this sample, almost all inappropriate placements were "test ups". The authors considered that the most important finding in this study was that the Placement Test could provide highly accurate information in only one direction. Given the base rate of approximately 75%, the Placement Test was quite effective in providing a threshold level below which it could be said with considerable certainty that the student was properly leveled, but the decision to "test up" when the threshold score was exceeded could not be made with similar accuracy. It remains to be seen whether this notion of the placement test as a threshold measure will be supported in a new sample that contains a larger portion of "test downs".

Kelley, P. R. Combined common person and common item equating of medical science examinations. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

The use of the Rasch model for equating medical examinations has passed through a three-phase evolution. First, was the equating of current and reference tests via a combined common person, common item procedure based on subtest calibrations of item difficulties. Next, current and reference tests were equated via a common item procedure based on total test calibrations of item difficulties. Finally, came the common item equating of a current examination, with difficulties calibrated on the total test, to a combination of several reference examinations, in other words, to the item bank.

This step of equating to the bank scale is important because it will allow for maximum flexibility in use of the item bank for test design and development. It will also provide a broader examinee base for the conversion of Rasch person abilities to scores on a traditional scale.

Once the item difficulties are in place on the bank scale, it will not be assumed that they can be used forever without being checked. Procedures to detect drift in item difficulty and fit will be implemented.

Use of the Rasch model in this way will enable the medical examiners to monitor, over time, comparability of examinee samples and the difficulty of items. This is expected to enhance their ability to monitor

the comparability over time of the standards set for passing the tests for purposes of certification.

Kiefer, E. & Bramble, W. The calibration of a criterion-referenced test. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1974.

The authors think that the crucial difference between criterion-referenced measurements and normative based testing is the way that scores are interpreted. The question is whether the test is designed for the purpose of comparing a score to some criterion or standard or whether it is designed to make comparisons between individuals.

The following questions are raised by the use of criterion-referenced measurements:

1) assuming that a standard or criterion has been chosen, how generalizable is it?

2) What is the relationship between the test items and test scores and those standards?

3) Given that the scores are compared to a standard, with how much precision can one state whether a particular score represents attainment or above the standard? That is, how good is the decision which classifies students as having either attained or not attained that standard?

It is with these questions in mind that the authors chose to calibrate a criterion-referenced test using the Rasch model.

The data for this investigation were item responses on an 84 item final exam from an introductory educational psychology course at the University of Kentucky. The subjects were 201 U. of K. undergraduates enrolled in the College of Education.

The items for the test were fitted to the Rasch model. The process by which the test was calibrated was based on the procedures outlined by Wright & Panchepakesan (1969); the six steps are outlined by the authors.

A major limitation of the procedure followed by the authors was the problem of arriving at a decision about what to do with items which measure important performances but do not fit a latent trait. For a Rasch model, there are 3 main reasons why an item does not fit a latent trait:

1) the item measures a different trait,

2) the item is poorly written and does not measure the desired performance,

3) the item measures the trait but does not fit the model because of the model's restrictive assumptions (i.e. that all the discrimination parameters are equal to one).

The authors advocate that the test maker should hypothesize which items measure which trait, calibrate the items, and look closely at items which do not fit the model to see if the items are good ones. If they are not good ones, re-write them, if they are good ones, keep them.

A second limitation of their procedure is the size of the sample needed to generate stable parameter estimates. Wright suggested to them that a group of at least 200 respondees is necessary. This could be a problem for a classroom teachers.

The authors conclude that the Rasch model, and other latent trait models, are potentially useful for the calibration of criterion-referenced tests. With a large pool of calibrated items, the tester can use any sub-set of them to estimate ability on the trait. This makes the comparison between the student's ability and the criterion a meaningful one without resorting to either a normative rubric or a defense of the criterion and the domain from which the items were selected. Given a criterion, it is possible to say, without regard to a sample of persons or the sample of items, what the probability of a person meeting the criterion is. In addition, given a large pool of items it is possible to estimate person ability to practically any degree of accuracy. This insures a more precise determination of the probability that the person has exceeded the criterion.

The authors are convinced then that the calibration of a criterion-referenced test leads to more consistency between the purposes of such testing and the interpretation of the scores. It leads to more generalizability about the meaning of the scores and more precision concerning the extent to which a score represents passing a criterion. Despite some limitations of this procedure it has great potential in its application.

Kingsbury, G. G. & Weiss, D. J. An adaptive testing strategy for mastery decisions. Research Report 79-5, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, MN. 55455, 1979.

In an attempt to increase the efficiency of mastery testing while maintaining a high level of confidence for each mastery decision, the authors applied the theory and technology of item characteristic curve (ICC) response theory and adaptive testing to the problem of judging individuals' competencies against a pre-specified mastery level to determine whether each individual is a "master" or a "nonmaster" of a specified content domain. Items from two conventionally administered classroom mastery tests administered in a military training environment were calibrated using the unidimensional three-parameter logistic ICC model. Then, using response data originally obtained from the conventional administration of the tests, a computerized adaptive mastery testing (AMT) strategy was applied in a real-data simulation.

The AMT procedure used ICC theory to transform the arbitrary "proportion correct" mastery level used in traditional mastery testing to the ICC achievement metric in order to allow the adaptation of the test to each trainee's achievement level estimate, which was calculated after each item response. Adaptive testing continued until the 95% Bayesian confidence interval around the trainee's achievement level estimate failed to contain the prespecified mastery level. At that point testing was terminated, and a mastery decision was made for the trainee.

Results obtained from the AMT procedure were compared to results obtained from the traditional mastery testing paradigm in terms of the

reduction in mean test length, information characteristics, and the correspondence between decisions made by the two procedures for three different mastery levels and for each of the two tests. The AMT procedure reduced the average test length 30% to 81% over all circumstances examined (with model test length reductions of up to 92%), while reaching the same decision as the conventional procedure for 96% of the trainees.

The authors note and discuss additional advantages and possible applications of AMT procedures in certain classroom situations; further research questions are suggested.

Kingsbury, G. G. & Weiss, D. J. Effect of point-in-time in instruction on the measurement of achievement. Research Report 79-4, Department of Psychology, University of Minnesota, Minneapolis, Mn., August, 1979.

The authors point out that item characteristic curve (ICC) theory has potential for solving some of the problems inherent in the pretest-test and test-posttest paradigms for measuring change in achievement levels. However, if achievement tests given at different points in the course of instruction tap different achievement dimensions, the use of ICC approaches and/or change scores from these tests is not desirable. This problem is investigated in two studies designed to determine whether or not achievement tests administered at different times during a sequence of instruction actually measure the same achievement dimensions.

To investigate possible changes in dimensionality between different points in instruction, aspects of the dimensionality of achievement test data were examined prior to instruction, at the peak of instruction, and up to a month following the peak of instruction. Data used were conventional and adaptive achievement test data administered to students in a general biology course at the University of Minnesota.

Results raised questions about the utility of the pretest-test paradigm for measuring change in achievement levels, since a comparison of ICC parameter estimates indicated that a change in the dimensionality of achievement had occurred within the short (4-week) period of instruction. This change was also observed using a factor analytic comparison.

Use of the test-posttest paradigm to measure retention was supported, since a regression comparison of students' achievement level estimates did not indicate any significant change in the achievement metric up to 1 month after the peak of instruction. The significance of this results for the use of adaptive testing technology in measuring achievement is described.

Implications of these studies and the use of ICC theory in the measurement of achievement, as well as some potential limitations in terms of generalizability of these results, are discussed.

Kingsbury, G. G. & Weiss, D. J. Relationships among achievement level estimates from three item characteristic curve scoring methods. Research report 79-3, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, Mn., 1979.

This study compared achievement level estimates from three item characteristic curve (ICC) scoring methods using the one-, two-, and three-parameter ICC models. The three scoring methods were maximum-likelihood normal, maximum-likelihood logistic, and Owen's Bayesian scoring method. Data included all possible response patterns from a hypothetical five-item test, as well as response patterns from live administration of a conventional and adaptive achievement test. For the conventional and adaptive test data, correlations among achievement level estimates were examined as a function of test length.

Results for all data sets showed a high degree of similarity among θ estimates for the one- and two-parameter data, with slight decreases in correlations as information on the discrimination parameter was used in scoring. When the third ("guessing") parameter was used in scoring the item response data, correlations among θ estimates were reduced, particularly for the adaptive test data. The data also showed an increasing tendency for the maximum-likelihood methods to result in convergence failures as the third parameter of the ICC was used in scoring. In general, the adaptive test data were less likely to result in convergence failures than were the conventional test data. The data also illustrated how each of the three scoring methods tend to utilize ICC parameter information in arriving at θ estimates and the relationships of these estimates to a number-correct scoring philosophy. The authors discuss the advantages and disadvantages of each of the scoring methods. They suggest that future research examine the relative validities of scoring methods and model combinations.

Klein, D. F. & Cleary, T. A. Platonic true scores: Further comment. Psychological Bulletin, 1969, 71, 278-280.

The true score in classical test theory is defined as an expected value. Some people have assumed incorrectly that this true score is also necessarily accurate. Unfortunately, the intuitively appealing accurate alternative to the classical true score, the "Platonic" true score, does not lead to the standard classical theory relationships among true, error, and observed scores. The present article attempts to clarify some questions regarding these two conceptions of true scores. Scales of measurement, types of distributions, and conditions where the two true scores are equivalent, and error reduction, are discussed.

This is Klein & Cleary's response to Levy's attack on their original article on Platonic true scores.

Klein, S. P. & Kosecoff, J. Issues and procedures in the development of criterion referenced tests. ERIC clearinghouse on tests, measurement and evaluation. Princeton, New Jersey: Educational Testing Service, 1973.

This paper has attempted to outline the basic steps and procedures in the development of criterion referenced tests as well as the issues and problems associated with these activities. In addition, representative CRT systems have been reviewed. From this analysis it is clear

that the developer of a CRT must answer a number of questions in order to clarify the nature and purpose of a CRT.

1) For what decision areas and purposes is the CRT most applicable?
2) What areas and objectives does the CRT cover and how were these objectives derived and organized?
3) How broadly or narrowly are the objectives defined?
4) How were the test items or tasks chosen to measure the objectives defined and developed?
5) How dependent are the items on particular instructional materials or programs? And what is their applicability to different kinds of students?
6) What methods were used to improve the items on the CRT and why were they chosen relative to the purpose of the instrument?
7) How was the validity of the CRT established?
8) What kinds of scores should be reported for a CRT and what is the justification for these scores, especially those involving "Mastery?"
9) How was the test finally put together, what compromises had to be made, and how were they resolved?
10) In what ways will packaging of the CRT facilitate its use?

The CRT systems reviewed in this paper are:

1) California Test Bureau-McGraw-Hill (CTB) - Prescriptive Mathematics Inventory (PMI).
2) Comprehensive Achievement Monitoring (CAM).
3) Individualized Criterion Referenced Testing (ICRT).
4) Instructional Objectives Exchange (IOX).
5) Minnemast curriculum Project-University of Minnesota.
6) National Assessment of Educational Progress (NAEP).
7) Southwest Regional Laboratory (SWRL).
8) System for Objectives Based Assessment - Reading (SOBAR); Center for the Study of Evaluation: UCLA
9) Zweig and Associates.

Koch, W. R. & Reckase, M. D. Problems in application of latent trait models to tailored testing. Research Report 79-1, Tailored Testing Research Laboratory, Educational Psychology Department, University of Missouri, Columbia, Mo. 65211.

Computerized tailored testing procedures have been successfully applied in the past to the measurement of aptitude or ability. The latent trait models employed in these procedures make the basic assumption that the underlying latent trait being measured is unidimensional. However, achievement tests are commonly found to measure several factors. The purpose of the present research was to study the effects of using tailored tests for achievement measurement, knowing that the unidimensionality assumption would be violated. Of equal importance to the study was a comparison of the one- and three-parameter logistic models to each other as well as to a traditional paper-and-pencil achievement test. A total of 110 undergraduate students enrolled in an introductory educational psychology and measurement course at the University of Missouri-Columbia served as examinees for the study. A counterbalanced test-retest design was employed in which there were two separate test

sessions one week apart for each examinee, with both the one- and three-parameter tests administered at each session. The tailored tests were administered on Applied Digital Data Systems Consul 980 cathode ray tube terminals which were connected to an IBM 370/168 computer through a timesharing system. Relative efficiency curves, test-retest reliability coefficients, goodness of fit of the models, descriptive statistics, content validity and the correlation of the tailored test ability estimates with the traditional course exam scores were used to compare the models. Item pools were constructed through the use of linking procedures to place item parameters from different test calibrations onto the same scale. During the tailored test, items were selected for administration based on the information function, and maximum likelihood ability estimation was employed. In addition, an attitude survey was administered after each test session to determine student attitudes toward the tailored tests. The results of the study indicated that neither tailored test procedure performed as well as the traditional course exam in terms of reliability. However, the three -parameter procedure had higher test information and better fit of observed responses to the model than the one-parameter procedure. Neither the one-parameter nor the three-parameter tailored tests yielded satisfactory content validity. The attitude scale results indicated generally favorable student attitudes toward tailored testing.

Koch, W. R. & Reckase, M. D. Problems in application of latent trait models to tailored testing. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, 1979.

The purpose of this paper was to describe some of the difficulties which became evident while conducting tailored testing research at the University of Missouri-Columbia. The authors first discussed the rationale behind tailored testing and some of its primary characteristics.

They conclude that the results of applying tailored testing procedures to the measurement of unidimensional vocabulary ability were generally satisfactory. However, tailored testing applied to multidimensional achievement measurement presented many difficulties. Some of the possible causes of these difficulties include the small sample sizes used to calibrate the tests resulting in unstable item parameter estimates; a compounding of the instability of the parameter estimates during linking procedures; the possibility that latent trait models may not be robust with respect to violation of the unidimensionality assumption by multi-content achievement tests; and the nonconvergence of some tailored tests when using maximum likelihood ability estimation.

The authors note that very little can be taken for granted in setting up tailored testing procedures. They conclude that a great deal more research must be conducted to determine optimal levels of the various components that control tailored testing procedures.

Kolen, M. J. & Whitney, D. R. Accuracy of estimating two parameter logistic latent trait parameters and implications for classroom

tests. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

The authors contend that the problem of possible low accuracy in estimation of the latent trait parameters due to small sample sizes must be considered before latent trait theory can be useful for classroom tests.

The purpose of this study was to assess, using computer generated data, the accuracy of estimation in Birnbaum's two parameter logistic model with varying small sample sizes. The maximum likelihood (ML) procedure for estimating the latent trait parameters was compared to a method in which the observed relative frequencies were smoothed using an isotonic regression method prior to applying the ML procedure. Comparisons of accuracy in estimation were based on mean squared error, variance and bias in estimating the parameters. In addition, the obtained average variances were compared to an approximation to the Rao-Cramer lower bound for the variance in estimating each of the two parameters.

The relative frequencies were randomly generated 100 times for each of the 45 possible combinations of a_g (slope parameter - item discrimination), b_g (location parameter - item difficulty) and N. The parameters (a_g and b_g) were estimated for each of the replications using both of the procedures. The individual ability parameters were fixed and assumed known for all replications.

Results: The regular ML method led to more accurate estimation of the slope parameter (a_g) whereas the isotonic method yielded more accurate estimation of the location parameter (b_g). The results indicate that since the isotonic method was not more accurate for both parameters and does not provide a reduction in computational labor, it probably would not be used exclusively in practice.

The results also indicated that the approximating variances approach closely the obtained variances for the larger sample sizes and at some combinations of the parameters for the smaller sample sizes.

In relation to classroom testing, unless samples of approximately 100 or more students are used, the latent trait parameter estimates will be fairly inaccurate (but the problem of inaccuracy in estimating the classical indexes also exists). For small samples, the estimates of the latent trait parameters are probably no less accurate than those for the classical indexes. The authors conclude that the choice between classical and latent trait theory should be made on the amount and quality of information each theory provides to the instructors.

Koslowsky, M. & Bailit, H. A measure of reliability using qualitative data. Educational and Psychological Measurement, 1975, 35, 843-846.

The authors note that in many types of research activities, it is necessary to obtain a reliability measure for qualitative or unordered data. The procedures that are presently available cannot handle such data using the classical reliability measures. Finn's (1970) method assumed internal type data, and Goodman and Kruskal's (1954) formula for

handling reliability of unordered data is good for only one item at a time. This paper expands the Goodman & Kruskal formula, and discusses an approach for calculating the inter-rater reliability for a series of items across many subjects. The procedure is considered to be analogous to the usual reliability determination for an achievement test or an attitude test.

Kriewall, T. E. Aspects and applications of criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972.

The measurement information generated by CRT's is designed for use in instructional management systems where classifications of pupils for treatment are to be decided on the basis of minimal data consistent with predetermined limits for the errors of misclassification. The measures obtained are content-specific estimates of proficiency useful for the stratification of learning groups on a day-to-day basis if need be. By sampling across items rather than across persons, absolute measures of proficiency are obtained which can be reliably interpreted for non-randomly selected pupils, the pupils of particular instructional concern. The model is designed for wide variety of applications but retains in the concept of proficiency a simple and useful index for instructional management. The empirical data generated have clear implications for instructional decision-making.

Some of the applications of item-sampling theory include the ability to;

1) categorize learners into temporary learning groups on the basis of a common requirement for instructional treatment (Diagnosis and Prescription Function);

2) assess the relative effectiveness of competing instructional treatments (Instructional Assessment Function);

3) to determine, in the case of established instructional segments having predetermined performance standards, which individuals have acquired minimal standards of proficiency required for mastery and which learners require further prescriptive assistance (Quality Control Function);

4) in the case of curriculum development, to indicate hierarchical relations within a content sequence (Curriculum Design Function).

Krippendorff, K. Estimating the reliability, systematic error and random error of internal data. Educational and Psychological Measurement, 1970, 30, 61-70.

The author states that the analyst of a recording instrument may wish to obtain the following:

1) An estimate of the reliability of a population of data over all observers in the universe using the recording instrument. This measure is called data reliability and can be interpreted as a measure of the confidence in data.

2) An estimate of the extent to which data reliability could be improved if scale values were to be transformed or their definitions

were to be modified for the individual observers. This measure assesses the systematic error of the recording process, which, together with a measure of the random error may be said to account for the lack of data reliability.

3) An estimate of the reliability associated with each individual observer, often called individual reliability. Such an estimate permits the identification of observers who are detrimental to achieving high data reliability. Deviant observers need either more instruction or cannot be employed in the process of collecting data.

4) An estimate of the extent to which each observer is corrigible by further instruction. Such an estimate would assess systematic observer biases which together with the individual's random error account for lack of individual reliability.

5) Finally, there is needed an indication of the extent to which a random sample of observers agree on the scoring of each unit of recording. This measurement may be called unit reliability and allows one to identify sources of unreliability within the set of observations.

In his article, the author aims at explicating these analytical devices.

Levy, P. Platonic true scores and rating scales: A case of uncorrelated definitions. Psychological Bulletin, 1969, 71, 276-277.

The demonstration by Klein and Cleary that the assumptions of classical test theory are untenable for Platonic true score theory is based upon a confusion of the nominal and interval scale meanings of the digits 1 and 0. Classical test theory applies, by definition, only when its definitions are allowed to operate. Their reference to the work of Sutcliffe did not support their case.

This is a rebuttal of Klein & Cleary's article on Platonic true score and errors in psychiatric rating scales in Psychological Bulletin, 1967, 68, 77-80.

Livingston, S. A. Reliability of tests used to make pass/fail decisions: Answering the right questions. Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, 1978.

The traditional reliability coefficient and standard error of measurement are not adequate measures of reliability for tests used to make pass/fail decisions. Answering the important reliability questions requires estimation of the joint distribution of the true and observed scores. Lord's "Method 20" estimates this distribution without the deficiencies of other methods. New output formats condense the estimated distribution into readily useable information, including a 2x2 contingency table, conditional true-score distributions, and an index of decision-making efficiency.

Livingston talks about "minimally acceptable performance" e.g. passing a final exam for a course, rather than "mastery-nonmastery" of the material. He asks three questions:

1) Of those persons who passed the test, how many would have passed if the test had been perfectly reliable? Of those who failed, how many would have failed if the test had been perfectly reliable?

2) What is the distribution of true scores in the group of persons who passed the test? What is the distribution of true scores in the group of persons who failed the test?

3) What is the decision-making efficiency of the test, as compared with that of a perfectly reliable test?

The author maintains that the traditional reliability coefficient and the standard error of measurement will not answer these questions but that Lord's "Method 20" will. Method 20 assumes that the conditional distribution of observed scores, for persons with a given true score, is a "compound binomial distribution" (this assumption is less restrictive than the assumption of a binomial distribution.)

It is noteworthy that for Livingston's data, sample size was 3,274 in which case the binomial and the compound binomial practically "converge".

Lloyd, B. H. A comparison of methods of vertical equating. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, 1979.

The purpose of this study was to explore a practical application of the Rasch model to the vertical equating of levels of a mathematics computation test. Three aspects of the problem were considered:

1) the consistency of calibration of ability estimates on the same test level for different levels of ability;

2) the adequacy of vertically equating adjacent test levels when estimates of parameters are obtained from two groups of comparable ability; and

3) comparison of the results of equating directly two nonadjacent test levels with the results of equating them by pairwise chaining through an intermediate level. The author thought that if the Rasch model was appropriate for vertical equating of this test, the calibrations should be consistent in determining ability estimates for the separate ability groups, and the equating of both adjacent and nonadjacent levels should be invariant with respect to groups.

The results from this study indicate that the use of latent trait methods in vertical equating should be approached with extreme caution. The author notes that this warning seems especially appropriate in the vertical equating of achievement tests where by necessity the content specifications of the test appreciably change with test level. She raises a note of caution for those test developers assembling achievement test item banks based on item calibration by latent trait methods.

Lord, F. M. An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989-1020.

The main purpose of this report is to point out certain kinds of information that are needed by anyone concerned with mental testing and to illustrate how such information can be provided by Birnbaum's methods, provided the underlying mathematical model is suitable for the data at hand. Lord hopes to achieve this purpose by presenting the results

obtained for one test: the College Entrance Examination Board's Verbal Scholastic Aptitude Test. The results represent a statistical analysis of the answer sheets of a sample of 2, 862 examinees who took the test on May 2, 1964.

The assumptions underlying Birnbaum's three-parameter logistic model may be stated loosely as follows:

1) The test items have only one psychological dimension in common.
2) The test items are scored either "right" or wrong".
3) The probability that an examinee will answer a given item correctly is a three-parameter logistic function of his verbal ability.

Lord notes that one can hardly expect the model to hold exactly; what is important is whether the model can provide trustworthy approximate answers to important practical questions.

It is noted that the results reported for the SAT are tentative and are given for illustrative purposes only.

Lord, F. M. Do tests of the same length have the same standard errors of measurement? Educational and Psychological Measurement, 1957, 17, 510-521.

Lord notes that from one point of view, tests of the same length have the same standard errors of measurement. Several questions regarding these standard errors are discussed in this paper and answers are given.

Lord, F. M. Estimating true-score distributions in psychological testing (an empirical Bayes estimation problem). Psychometrika, 1969, 34, 259-299.

The following problem is considered: Given that the frequency distribution of the errors of measurement is known, determine or estimate the distribution of true scores from the distribution of observed scores for a group of examinees. Typically this problem does not have a unique solution. However, if the true-score distribution is "smooth," then any two smooth solutions to the problem will differ little from each other. Methods for finding smooth solutions are developed:

a) for a population and
b) for a sample of examinees.

The results of a number of tryouts on actual test data are summarized.

This is the article about "Method 20" which Livingston refers to in his '78 NCME paper.

Lord, F. M. A note on the normal ogive or logistic curve in item analysis. Psychometrika, 1965, 30, 371-372.

Lord notes that it is common to assume that the proportion of correct answers to an item has a normal-ogive or logistic relationship to the total test score. He shows that this is a mistaken and an undesirable notion.

Lord, F. M. A significance test for the hypothesis that two variables measure the same trait except for errors of measurement. Psychometrika, 1957, 22, 207-220.

Lord derives the likelihood-ratio significance test for the hypothesis that after correction for attenuation two variables have a perfect correlation in the population from which the sample is drawn.

Macready, G. B. & Dayton, C. M. A two-stage conditional estimation procedure for unrestricted latent class models. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

The purpose of this paper was to describe a two stage conditional estimation procedure which results in reasonable estimates of specific models even though they may be non-identifiable. This procedure involves the following stages:

1) establishment of initial parameter estimates and
2) step-wise maximum likelihood solutions for latent class probabilities and classification errors with iteration of this process until stable parameter estimates across successive iterations are obtained.

Macready, G. B. & Merwin, J. C. Homogeneity within item forms in domain referenced testing. Educational and Psychological Measurement, 1973, 33, 351-360.

This paper considers the nature of the relationships among items within item forms and how these relationships compare with an ideal case for diagnostic tests in which if a person gets one item within an item form right, then he should get all items within the item form correct.

The results suggest that, in most cases, item forms which generate items of moderate difficulty can be used to obtain relatively homogeneous sets of items of equivalent difficulty for a defined population of subjects. Such item forms provide sets of items superior to those which would be expected if item difficulties alone were used to group items into sets. This suggests that the means used in defining the replacement-set structures by attempting to objectify the intuitive categories ordinarily used by teachers in constructing diagnostic tests is at least a reasonable first effort.

The results also suggest a basis for identification of item forms which will generate homogeneous items of similar difficulty. Using this information, it is possible to determine whether the breadth of an item form is appropriate and if not, identify changes which will lead to an item form of more useful breadth.

McLarty, J. R. Multi-level item analysis. Paper presented at the annual conference of the California Society of Educational Program Auditors and Evaluators, San Francisco, 1979.

The author notes that most educational data are multi-level; data are collected from individuals (students, teachers, etc.) within groups (classrooms, schools). Observed group effects may be the result of the ways in which individuals are selected into the groups or of the effects on individuals of the programs and processes which take place within the

groups. Empirical item analysis usually combined information from students in multiple groups, resulting in reliability and validity coefficients which combine between and within group relationships. This paper presents an application of multi-level regression techniques to item analysis using Biology test data for the U.S.A. from the IEA six-subject survey. Statistical procedures for disaggregating between and within group relationships are presented, and possible interpretations of the results and implications for program evaluation are discussed.

Marco, G. L., Peterson, N. W. & Stewart, E. E. A test of the adequacy of curvilinear score equating models. Paper presented at the Computerized Adaptive Testing Conference, Minneapolis, 1979.

This study is the first part of a fuller study, the purpose of which is to examine the adequacy of score equating models when certain sample and test characteristics are systematically varied. The emphasis in this part of the study is on curvilinear models. The second part focuses on linear models. This study is more comprehensive than previous studies of equating models in that it includes a greater variety of equipercentile, linear and ICC models and investigates equatings based on dissimilar samples as on random samples.

Results are presented in tables and figures. Following an explanation of the tables and figures, salient features of the results are discussed, some limitations on their interpretation are suggested and several conclusions are presented.

Maxwell, A. E. The effect of correlated errors on estimates of reliability coefficients. Educational and Psychological Measurement, 1968, 28, 803-811.

The procedure whereby the "reliability" coefficient of a test can be derived by analysis of variance is reviewed. The assumptions underlying the analysis of variance model are noted and it is shown that if the error terms in the model are not independent then the estimate of the reliability coefficient will be biased, and in most commonly occurring cases will be an overestimate.

Maxwell, A. E. Maximum likelihood estimates of item parameters using the logistic function. Psychometrika, 1959, 24, 221-227.

The logistic function is proposed as an alternative to the integrated normal function when estimating parameters of test items. The logistic curve is described; an iterative method for finding maximum likelihood estimates of its parameters is given, and an example of its use is presented.

Mehrens, W. A. & Ebel, R. L. Some comments on criterion-referenced and norm-referenced achievements tests. NCME measurement in education, 1979, 10, No. 1, 1-7.

This paper is divided into two sections:

1) the controversy and problems associated with defining tests (i.e. criterion-referenced tests and norm-referenced tests); and

2) the problems associated with using tests (i.e. standardized versus tailor-made achievement tests, uses for criterion-reference interpretation, mastery testing, and uses for norm-reference interpretations).

The authors conclude that there is a place in educational measurement for both norm-referenced and criterion-referenced test interpretations. The question, they state, is not which interpretation to use, but when to use each. They stress that local tailor-made tests are desirable supplements to external standardized tests, not superior alternatives.

Merz, W. R. & Rudner, L. M. Bias in testing: A presentation of selected methods. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

Selected methods for examining the test performance of members of identifiable groups for fairness were presented. Two sets of methodologies were identified: one in which an intact test is administered to members of different groups to provide data for selection; the other, in which items from a pool are examined for systematic differentiation among groups. The purpose of this paper was simply to describe the methods. No attempt to evaluate them was made.

Under the first condition of administering an intact test to members of different groups, seven approaches to regression analysis were reviewed. All seven attempt to predict from a selection or placement instrument to a criterion of success. The first method, labeled the regression model by Petersen & Novick defines a test as fair if there are no consistent non-zero errors of prediction for members of each subgroup of the population. The second method described by Thorndike was called the constant ratio model by Petersen & Novick. A test is fair if it identifies applicants for selection in such a way that the ratio of the proportion selected to the proportion successful is the same in all subpopulations. A third approach was proposed by Einhorn & Bass and labeled the equal risk model by Petersen and Novick. It defines a test as fair when all persons selected are predicted to be above a specific minimum point on the criterion with a specified degree of confidence.The fourth method was proposed by Darlington and suggests a model which would replace the concept of cultural fairness with another which he labels cultural optimality; hence, it was called the culture modified criterion approach by Petersen & Novick. Cole proposed a fifth method labeled the conditional probability model by Petersen & Novick. In this model, a test is regarded fair if, given satisfactory criterion performance, individuals have the same probability of selection regardless of group membership. The sixth model was proposed by Linn and defines a test as fair if all applicants who are selected are guaranteed an equal, or fair, chance of being successful regardless of group membership. This model was labeled the equal probability model by Petersen & Novick. The seventh model to be reviewed was proposed by Gross & Su and was labeled the threshold utility model by Petersen & Novick. It states that a test is fair if an individual from a subpopulation is selected

when his/her predicted score reaches a specific minimum point on the criterion which has been modified in such a way that the expected utility of the selection process is a maximum.

Under the second condition in which items from a pool are examined for systematic differentiation among groups, six approaches were reviewed. They were:

1) Analysis of variance;
2) Transform Item Difficulties;
3) Correlation approaches;
4) Factor Analytic Approaches;
5) Distractor Response Analysis;
6) Item Characteristic Curve Theory Approaches.

Merz, W. R. & Grossen, N. E. An empirical investigation of six methods for examining test item bias. Report No. NIE-6-78-0067. National Institute of Education, Washington, D. C.

The purpose of this investigation was to examine six approaches for assessing test item bias. The methods employed were Transformed Item Difficulty, Point Biserial Correlations, Chi-Square, Factor Analysis, One Parameter Item Characteristic Curve and Three Parameter Item Characteristic curve. Data sets for analysis were generated by a Monte Carlo technique based on the three parameter model; thus four parameters were controlled: total score distributions, item difficulties, item discriminations and guessing. Only the difficulty parameter was biased. Results indicated that Transformed Item Difficulty had highest correlations with generated bias. The Three Parameter Item Characteristic Curve and the One Parameter Item Characteristic Curve were next highest. Factor Analysis then Chi-square followed next. Point Biserial correlation functioned erratically. Results of the analysis are compared and recommendations on the use of each method are presented.

Meskauskas, J. A. Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. Review of Educational Research, 1976, 46, 133-158.

The purpose of this paper is to investigate the mastery models underpinning the techniques being proposed and to review the procedures suggested for setting the pass-fail point.

Mastery models fall into two broad categories:

1) Continuum models - mastery as an area on a continuum
2) State models - mastery as all or none.

Characteristics common to continuum models are:

1) Mastery is viewed as a continuously-distributed ability or set of abilities.

2) An area is identified at the upper end of this continuum, and if an individual equals or exceeds the lower bound of this area, he is termed a master.

3) The goal of measurement is to obtain information for the purposes of educational decision-making, which explicitly follows the classification decision.

The author reviews Nedelsky's minimum pass level (MPL) method; Ebel's method of passing score estimation; the Kriewall binomial-based model;

Characteristics common to state models are:

1) Criterion-referenced test (CRT) true-score performance is viewed as an all-or-none dichotomous task.

2) The standard or cutting score that should be used in an error-free situation is implied as part of the model.

3) Considerations of measurement error essentially always result in the adoption of standards that demand less than the model seeks.

The author reviews Emricks' mastery testing evaluation model, Roudabush's dichotomous true-score models.

Other models which may be considered mixed include Millman's binomial-based decision model; the Davis and Diamound Bayesian method; the work of Novick and collaborators.

This paper is organized around the views of various authors regarding the nature of mastery and the way it is acquired since differences in conceptualization result in very different approaches to evaluation.

Meskauskas, J. A. Standard setting procedures for medical specialty boards. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

The purposes of this paper are:

1) to describe the ways in which standard-setting procedures are carried out in the medical specialty certification arena,

2) to briefly discuss some experiments which have been conducted with alternatives to currently-used normative standards, and

3) to speculate on some of the reasons why these experiments have had limited success.

The defensibility of normative as well as absolute standards is discussed.

Meskauskas' view is that the only philosophically defensible standards are those which judge the adequacy of a performance solely on the basis of how well an examinee deals with the content of the examination. (Since judgment is needed to obtain these standards, to call them "absolute" connotes too much -hence he uses the term "content-referenced").

In summary, the experiences to date in this medical area do not support a hope that one can successfully apply an existing model or technique to this area. However, there is a growing support for content-referenced standards and unflagging philosophical support for the importance of the effort. It is believed that these circumstances, when combined with a well-thought out research effort in the decision processes of standard-setting, will undoubtedly lead to standards which are defensible and supportable by all parties-decision-makers, public and candidates.

Millman, J. Criterion-referenced measurement. Prepublication draft, 1974.

The purpose of this monograph is to acquaint the reader with the present state of the art of criterion-referenced measurement and to suggest directions that further inquiry and the future literature might take.

The monograph covers the following topics:

1) A concept of criterion-referenced tests
   a) Traditional definition
   b) A refined view including domain-referenced tests and differential assessment devices.
   c) the relation of test uses to desirable test characteristics.
   d) some other terminology for tests including content standard tests, objective-based tests and mastery tests.
2) Domain-referenced tests
   a) defining the item population including linguistically-based schemes, item forms, facet analysis, amplified objectives, domain size and the defining facets.
   b) generating and selecting test items.
   c) establishing a passing standard
   d) determining test length including classical binomial model and a Bayesian model.
   e) evaluating the domain definition and test including reliability and validity.
3) Differential assessment devices
   a) criterion groups observed including choosing the criterion variable, developing an item pool, selecting the specific items, choosing procedures for validation.
   b) criterion groups unobserved.
4) Educational applications of domain-referenced tests
   a) needs assessment
   b) individualized instruction including domain-referenced tests, differential assessment devices.
   c) program evaluation
   d) teaching improvement and personnel evaluation.

Morgan, P., Kosecoff, J. Walker, C. and Keesling, J. W. It's the metric that counts or criterion-referenced schizophrenia. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

The purpose of this paper is to lend support to the view that different decision purposes require Criterion-referenced tests (CRTs) with different physical and theoretical properties and that therefore, CRTs must be evaluated in terms of the context in which they will be used. To do this, two situations in which CRTs are being used, the classroom resource and the program evaluation context, were considered and a sample of criteria that have been used or are being used to evaluate CRTs were assigned weights appropriate to each context. Then, 28 commercially published CRTs were rated using each set of weights.

The dually-weighted set of fourteen criteria are:

1) Overlap of objectives across grade levels
2) Grade level coverage
3) Number of test forms
4) Special equipment needed for test administration
5a) Average time for administering a single test
5b) Average time for administering all tests
6) Machine/self-scoring
7) Score interpretation
8) Curriculum match
9) Who can interpret scores
10) Formal field test
11) Stability/number of items per objective
12) Sensitivity to instruction
13) Subject area comprehensiveness
14) Availability of comparative information.

Each CRT was independently reviewed twice using the 14 criteria and discrepancies were resolved by the reviewers. When sufficient information could not be obtained to rate a criterion, a score of zero was assigned. Review scores were computed as the proportion of possible points earned by a CRT and separate scores for the two weighting systems were reported for each CRT.

There is considerable variation in the scores each CRT earned using the two different weighting scales. When rated within the classroom resource context, the highest percentage obtained was 80 percent, with a low of 43 percent. Within the program evaluation context, the spread in the range of high and low percentages was quite similar, being 83% and 43%, respectively. A tentative conclusion, based on a review of the percentages earned by a CRT in these two contexts, might be that in trying to achieve both the classroom resource and program evaluation functions, CRTs are, for the most part, only marginally fulfilling each purpose.

The findings of this study support the view that the same criteria cannot be used for all purposes and that therefore, a CRT must be developed, validated and evaluated in terms of the purpose for which it is intended.

Moy, M.L.Y. & Barcikowski, R. S. Item sampling: Optimal number of people and items. The Journal of Experimental Education, 1974, 42,

Using a computer-based Monte Carlo approach to generate item responses, the results of this study indicate that, when item discrimination indices are considered, item sampling procedures having the same number of observations have different standard errors in estimating both test mean and test variance. With certain types of tests, a single item sampling plan would not yield optimal, i.e., smallest standard error, estimates of both μ and σ^2. That is, one sampling plan would be needed to optimally estimate μ and another to optimally estimate σ^2. In addition, it was found that single exhaustion of the item set was sufficient for estimating both μ and σ^2.

Moyer, J. E. Alternative reliability indices. Unpublished paper. Michigan Department of Education, 1979.

In an effort to find the appropriate reliability estimates for criterion-referenced tests, Michigan Educational Assessment Program (MEAP) has utilized several methods, both traditional and new, to document the reliability of its tests. The purpose of this paper is to present the results of those efforts.

The first study compares kappa and Kuder-Richardson Formula 20 as reliability estimates. The second study presents some preliminary results of a test-retest reliability study which used kappa as one of the correlation coefficients. In the third section, information is presented comparing kappa, as an item discrimination index, to the Brennan (B) index and phi.

The most important results of these attempts to show the reliability of the MEAP tests are:

1) kappa is dependent on variance for its computation.
2) kappa and phi behave in almost the same fashion.

Moyer, J. E. & Fishbein, R. L. A comparison of Kuder-Richardson formula 20 and Kappa as estimates of the reliability of criterion-referenced tests. Unpublished paper, Michigan Department of Education, 1979.

This paper attempted to provide some answers to the question: If traditional and new methods were used with data obtained from administering criterion-referenced tests, would different decisions about the reliability of those tests be made on the basis of the two reliability estimates, Kuder-Richardson Formula 20 and Kappa?

The results of the ANOVA showed that the differences between the mean KR-20's for the three groups (parallel forms, KR-20, and Kappa) were not found to be significant at the .05 level. Further results showed that there was a difference of the mean kappas for the three groups at the .05 level of significance.

The authors recommend that further research needs to be done using the KR-20 coefficient computed across the items of two or more randomly parallel tests and the kappa computed for those tests, in order to determine the relationship between KR-20 and Kappa. Also, the behavior of KR-20 and Kappa under criterion-referenced test conditions needs to be researched.

Nitko, A. J. & Feldt, L. S. A note on the effect of item difficulty distributions on the sampling distribution of KR-20. American Educational Research Journal, 1969, 6, 433-437.

To investigate the possibility of the effect of the item difficulty distribution on the sampling distribution of KR-20, two extreme types of distributions of ϕ_j's were constructed:

1) a uniform distribution over the range ϕ_j, = .20(.05).80

2) a highly concentrated distribution in the neighborhood of ϕ_j = .50 with the range ϕ_j = .45(.05).60.

Several Monte Carlo experiments were conducted to examine the effect of these item difficulty distributions on the sampling distribution of KR-20. Ten distributions of KR-20 were obtained-two under each of five levels of population reliability.

Data suggests that the effect of these two extremes of item difficulty distributions is minimal. Some of the small differences in percentiles which exist is attributed to sampling error and to the fact that the population reliabilities were not precisely equal for the two tests.

The authors contend that the data provide strong evidence that the form of the distribution of item difficulties has little effect on the sampling distribution of KR-20.

Oosterhof, A. C. Stability of various item discrimination indices. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, 1973.

This study was concerned with the degree to which various discrimination indices remain invariant when the item with which they are associated assumes membership within different sets of items. The normal procedure used in this tryout of items is to divide the total sample of items into several test forms and administer each form to a different group of subjects. One of the inherent assumptions involved in this procedure is that, as an item assumes membership within varying sets of items, the respective parameters associated with the item remain constant. If this assumption of stability is not true for a particular index of discrimination, then the items selected for inclusion in the final test form on the basis of such an index will be determined by the chance combination of items with which the item was originally associated, and not necessarily by the discriminatory power intrinsic to the item were it to be included in the final form of the test instrument.

The item discrimination indices compared in this study include point biserial coefficients, biserial coefficients, phi coefficients, estimates of tetrachoric coefficients, Gulliksen's item reliability, Findley's difference index, Flanagan's approximation and Davis' transformation of the product-moment correlation coefficient. An adaptation of the point-biserial coefficient using Livingston's (1972) criterion-referenced applications of classical test theory was obtained by substituting various criterion scores for the test mean.

The 50 item Verbal Reasoning subtest of the Differential Aptitude Tests was used as the source of items for this study and was administered to 10th grade students in 11 school districts in the state of Kansas.

Data from the study indicate that as an item is given membership within vaying sets of other items, Gulliksen's item reliability index and Findley's difference index are the most stable of the 24 discrimination indices investigated with respect to the invariance characteristics. A detailed discussion of the findings is reported.

Osburn, H. G. The effect of item stratification on errors of measurement. Educational and Psychological Measurement, 1969, 29, 295-301.

This paper shows that, in the case of matched item tests, the reduction in errors of measurement for tests constructed by stratified sampling as compared with tests constructed by random sampling from an infinite population of items, is a simple function of the variance of the difference between pairs of strata true scores. For unmatched item tests, the reduction in errors of measurement due to stratification is a function of the variance (across strata) of the strata mean true scores plus the variance of the difference between pairs of strata true scores.

These results predict that, in the case of matched item tests the largest reductions in errors of measurement will result from stratification on item content rather than item difficulty while for unmatched item tests just the opposite is true.

Osburn, H. G. Item sampling for achievement testing. Educational and Psychological Measurement, 1968, 28, 95-104.

This paper concerns the explicit definition of the universe of content, and the stratified random sampling of items from the universe of content so defined.

A universe defined test is a test constructed and administered in such a way that an examinee's score on the test provides an unbiased estimate of his score on some explicitly defined universe of item content. Two requirements for test construction are:

1) all items that could possibly appear in the test should be specified in advance,

2) the items in a particular test should be selected by random sampling or stratified random sampling from the universe of content.

Osburn's approach to defining a universe of content is to analyze the content area into a hierarchical arrangement of item forms and to develop a program for a digital computer that would compose item sentences given a suitable vocabulary and structural codes for the item forms.

An item form has the following characteristics:

1) it generates items with a fixed syntatical structure,

2) it contains one or more variable elements,

3) it defines a class of item sentences by specifying the replacement sets for the variable elements.

The principal advantage of item forms analysis is that it seems possible to characterize the universe of content as an abstract system while maintaining an unambiguous link between the system and the actual items that appear on any form of the test.

Osburn's treatment of item forms analysis draws on the basic features of Hively's approach with more emphasis on the hierarchical arrangement of item forms into a generalized system. The method is the reverse of Gagné's task analysis in that it proceeds from the general to the specific with an emphasis on the abstract system rather than on specific task elements.

The author contends that the abstract system together with the item generating program satisfies the properties of a universe defined test.

It is the author's stance that the mental test model that has been so successful in aptitude testing is not appropriate for across the board application to achievement testing.

Theoretical implications of a universe defined test included:

1) Reliability Theory:

In a universe defined test, a particular test or item sample becomes relatively unimportant and interest is focused on the universe of content. Concern is not focused on a specific test but rather with a procedure for estimating an individual's true score on a universe of content.

Classical mental test theory involving assumptions of test equivalence has been divorced from test content and true score content has been little more than a statistical fiction. The theory of generalizability (Cronbach et al) which assumes random or stratified random sampling of test conditions as a starting point has linked reliability theory with test content. Universe defined tests can be made to satisfy rigorously the assumptions of generalizability theory and constitute a practical means for implementation of the theory. In generalizability theory, the concept of the universe true score becomes meaningful.

2) Validity Theory:

For a universe defined test, what the test is measuring is operationally defined by the universe of content as embodied in the item generating rules. It should be possible to keep separate the concept of what a test is measuring from the concept of the extent to which the responses of a person sample to the universe of content are related to their responses to other classes of stimuli (construct validity, etc.). In response to Ebel's (American Psychologist, 1961, 16, 640-647.) plea for the operational definition of measurement procedures, Osburn suggests that the most important requirement for the operational definition of a test is the specification of the universe of content.

3) Item Analysis:

The author suggests that item analysis techniques be redefined if we are to preserve the idea of random sampling from a specified universe of content. Any decision to exclude items based upon item analysis data must result in a redefinition of the universe of content.

4) Normative Data:

The percent correct score of a universe defined test is meaningful (Ebel points out that to be meaningful, any test score must be related to test content as well as to scores of other examinees) because it is related to test content. On most psychological tests, a percent correct score is meaningless because the universe of content is not completely specified and random sampling is neglected.

5) Matched vs. Unmatched Data:

Matched-the same sample of items is administered to each subject in the person sample.

Unmatched-the items are randomly sampled for each subject. If the investigator is interested in the absolute score of an individual, it does not matter whether or not data are matched or unmatched. If he wants relative scores for individuals, matched data are required. If he wishes to estimate the mean for a group of persons, and proposes to generalize over persons and items, unmatched are preferred.

Patience, W. M. & Reckase, M. D. Operational characteristics of a one-parameter tailored testing procedure. Research report 79-2, Tailored Testing Research Laboratory, University of Missouri, Columbia, MO. 65211.

The primary objective of this investigation was to determine the effects of varying the program parameters, stepsize and acceptance range, as well as the item pool attributes, size and shape, on the bias and standard error of the maximum likelihood ability estimates obtained from tailored tests. Two main research questions were addressed. First, what value of stepsize and acceptance range provided the least bias and smallest standard error of ability estimates? Second, what shape and size of item difficulty distribution provided the least bias and standard error of ability estimates across the range of the latent trait?

Two programs, TREE1P and SIM1P, were used for investigating the effects of program parameters and item pool attributes. Results suggested that each of the variables played a role in affecting the magnitude of statistical bias and standard error at various points along the ability continuum. The results were presented as a guide for those involved in setting up a tailored testing procedure. Figures and tables are provided to facilitate applications of tailored testing procedures such that a minimum of bias and standard error of ability estimates could be attained.

Patience, W. M. & Reckase, M. D. Operational characteristics of a Rasch model tailored testing procedure when program parameters and item pool attributes are varied. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, 1979.

The primary purpose of the research described in this paper was to determine the operational characteristics of a one-parameter tailored testing procedure when program parameters and item pool attributes were varied.

Two FORTRAN programs were used for investigating effects of program parameters and item pool attributes. The input variables for both programs included:

a) acceptance range,
b) stepsize,
c) item pool size,
d) item difficulty values for the various sizes and shapes of item pools,
e) the true abilities for which an estimate was to be made utilizing the program parameters and item pool provided.

The results of this study were drawn from tables which summarized the results of the TREE1P and SIM1P programs. The tables are provided in this paper.

Plake, B. S. & Hoover, H. D. A methodology for identifying biased achievement test items that removes the confounding in an item-by-groups interaction due to possible group differences in instructional level. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

The paper presents a method for identifying biased achievement test items. One technique of screening a test for biased item is an analysis

of variance procedure. In studies using ANOVA procedures, the significant items-by-groups interaction is considered to be a signal that biased items may be present.

One reason for checking an achievement test for biased items is to remove items that are systematically favoring one group of examinees differently than would be expected by initial achievement level.

The hypothesis of interest is: "there is no interaction between items and groups." When a significant items by groups interaction is found, a follow-up procedure is recommended to identify the items contributing to the items by group interaction. Plake & Hoover recommend something called the Bonferroni follow-up procedure which is based on the interaction contrast.

Poggio, J. P., Asmus, E. P. & Levy, J. An investigation of responses to omitted items under formula scoring instructions. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

The study was designed as an empirical test of Frederic Lord's hypothesis ('75) that responses to omitted test items obtained from formula-scoring testing are random guesses. Two hundred twenty-one individuals participated in the study. Each participant was administered a course pre-test on which they were told to respond only to items that they felt they could make a knowledgeable response. Students were then later given an opportunity to respond to those items they originally omitted. Results of analyses on the omitted items found that:

1) student responses to these items exceed that expected by chance;

2) that the rank ordering of items based on item difficulties obtained from first attempt and second attempt responses remained greatly intact; and

3) factor analyses of item intercorrelations among first attempt and second attempt items resulted in assessment of a similar trait on both measures. The pattern of findings that the authors have observed fail to support the assumption that responses to omitted items are random. This conclusion is justified only within the testing conditions studied in this investigation.

In presenting his position Lord states that students perhaps need formal instruction in following test instructions. The authors found that the item mean over those items first omitted was .332. Their formula-scoring instructions encouraged students to respond only if one alternative could be omitted when judging each of the four alternatives presented with each item. For the situation examined in this study, the nature of such instruction should be to encourage students to attempt items and not to discourage them.

The authors examination of formula-scores and number-right scores as predictors of course achievement found the number-right estimators as superior. For the situation they studied they failed to find evidence to support Lord's position that the number-right score is an inadmissable estimator.

Popham, W. J. Technical travails of developing criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 1974.

The purpose of this paper was to recount the decisions, errors, and insights made during the 10X criterion-referenced test development enterprise. Major problems encountered come under the following headings:

1) An optimal number of tests. The question of content or skill generality manifests itself in connection with determining how many criterion-referenced tests to produce.

2) Choosing a domain. The following six considerations were used in deciding on domains:

a) general acceptance to teachers, subject matter specialists, the public;
b) transferability within the domain;
c) transferability outside the domain;
d) terminality;
e) amenability to instruction;
f) ease of scorability.

3) Domain homogeneity. The author reports that his staff for the 10X project was unable to solve this problem.

4) Cost and conscience. The author doubts that a commercial or nonprofit test development agency can engage in the development of truly high quality criterion-referenced tests without substantial external subsidies.

5) A technical wasteland. The author concludes that what is needed is a well financed, governmentally-initiated project to expand the present weak technological base for criterion-referenced measurement.

Popp, J. A. Toward a quantitative estimate of internal validity. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

The notion of internal validity is examined for purposes of establishing a method of quantitatively estimating it. The logical problem of identifying independent alternative hypotheses is considered as well as their relative plausibility weighting. Finally, the question of internal validity is viewed as being a form of Harmon's "inference to the best explanation".

The question of internal validity is taken to be: Did the treatment make any difference or have any effect upon the criterion variable? Internal validity is a matter of how well a particular instance of data collection or generation can be described and explained. The author states that within the confines of a single experiment, i.e., where there is no concern for generalizability, the only rationally defensible way of inferring "what happened" is the method described by Harman.

The reference to Harman is: G. H. Harman, The inference to the best explanation, Philosophical Review, 1965, 74, 88-95.

Ramsay, J. O. True score theory: A paradox. Educational and Psychological Measurement, 1971, 31, 715-719.

In classical mental test theory, if there is no a priori reason for accepting the statement, "there is no platonic true score," then it is usually not unreasonable to define true score as the expected value of observed score. Ramsay attempts to show that there are consequences of this assumption.

Reliability is defined as $\rho = \sigma_t^2 / (\sigma_t^2 + \sigma_e^2)$ where $\sigma_e^2 = \sigma_x^2 - \sigma_t^2$. By using a fundamental theorem about variance and noting that $E(x|t) = t$, the variance of observed score for a particular true score is limited. In order to see how different from zero this lower limit on reliability may be in practice, Ramsay proposes the beta distribution be used as a model for the distribution of true score.

The result is a "realistic" lower bound on reliability as a function of true score mean and variance. An example is given which expresses the lower bound on reliability as a function of true score standard deviation.

The author notes three ways out of this paradox:

1) Work only with scores transformed so as to be distributed on an infinite interval. This, he points out, seems to make the concept of true score even more artificial than when it was defined to be the expected observed score.

2) Replace this assumption with some new ones. The danger here is that the resulting test score theory will be stronger and contain even more parameters than can be handled computationally and theoretically.

3) Abandon the whole enterprise of describing test score behavior out of a predictive context and rely on standard statistical methodology to relate one test to another. The author notes that this is a radical approach which few may favor.

Reckase, M. D. A comparison of the one- and three-parameter logistic models for item calibration. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

The author notes that there has been an ongoing debate concerning the relative merits of the one- and three-parameter models, and that most of the research used simulated test items. The purpose of this paper was to extend the comparison of these two models to real data with reasonable sample sizes and to evaluate the models on both theoretical and practical grounds.

Five specific comparisons were made:

a) the evaluation of the goodness of fit of each of the models to the item response data;

b) the determination of the relationship between the ability estimates and the item responses;

c) the determination of the predictive validity of the ability estimates from the models in some limited cases;

d) the estimation of the minimum sample size required for each model to calibrate tests;

e) the determination of the relationship between the ability estimates obtained from the two models.

Sixteen data sets were used. The first eight were obtained from the administration of the Missouri School & College Ability Test (MSCAT), a standardized test, to groups of students throughout the state of Missouri. The second set was obtained from the administration of four classroom exams given to undergraduate students in a large measurement course. (Eight simulated test data-sets were also produced; these were generated to match various factor matrices using the usual linear factor analysis model).

Summary of Results:

a) the three-parameter model fit the test data better in all cases than the one-parameter model and there was a trend in the fit related to the dimensionality of the test;

b) the one-parameter model ability estimates shared more variance with the item responses than the three-parameter model;

c) there was no difference in the concurrent validity for small samples using the two models predicting classroom achievement tests;

d) the one-parameter model required smaller samples for calibration than the three-parameter model;

e) the ability estimates from the two models correlated highly for most of the data-sets.

The one-parameter model was preferred for use with small sample group data to predict outside criterion variables.

Reckase, M. D. Item pool construction for use with latent trait models. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

The purpose of this paper was to analyze six three-parameter logistic linking techniques and one one-parameter logistic linking technique to determine:

a) the sample size required to adequately link tests;

b) the number of overlapping items needed to link tests;

c) whether parameter estimate drift occurs when new tests are linked to an existing pool;

d) which procedure was best for linking item parameter estimates obtained from the three parameter logistic model.

The one-parameter linking did well regardless of the number of items in common between tests if sample sizes of 300 or more were used. The three-parameter linkings required a substantially larger sample size for stable linkings. A sample of 1000 was recommended. No clearly superior three-parameter linking procedure was found, although the major axis method applied to LOGIST calibrations was less accurate than the others. At the recommended 1000 sample size, the major axis ANCILLES method and maximum-likelihood LOGIST method seem to give a good combination of accuracy for estimation of the three-parameters. All of these conclusions were determined based upon multivariate achievement test data. If more unidimensional data were used, larger correlations would be expected.

Reichman, S. L. & Oosterhof, A. C. Strategy guidelines for the construction of mastery tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

In this paper, the authors propose a comprehensive model which would facilitate the instructional designer or individual instructor in concurrently considering various components which affect the construction of criterion-referenced tests used to make pass/no-pass decisions in regard to specified decision points. Components within the model include the average response latency time associated with the specific item format, the total amount of student time to be allocated to testing, the number of items selected for determining pass/no-pass decisions on each decision point, the number of mastery-status decisions made within the course, and the probability for individuals performing at a specified true domain score of being placed in the correct mastery state.

In order to demonstrate the empirical determination of probabilities of correctly classifying individuals into mastery/non-mastery categories, a 56 item test covering a well-defined domain of science information was administered to 1281 high school students. Baseline probabilities were determined under various domain and test sizes, using six different criterion levels in each case. Binomial expansions were then used to determine probabilities as test lengths were increased.

The data collected suggested that when a well-defined domain is established, as the actual domain size was reduced the average probability of misclassifying individuals at specific domain levels remained fairly consistent. Further, as the actual test size was reduced the baseline probability of misclassifying an individual at a given domain level also remained fairly consistent, while the standard deviation increased slightly. Increasing the criterion level resulted in an increase in the probability of misclassifying individuals with domain scores below that criterion level.

Application of the model demonstrated how the designer or instructor could manipulate components within the model in order to select the most efficient combination of factors to meet present needs. It was also demonstrated how this model could be used to make testing decisions for specific types of students based upon estimates of student performance with the content domain. The need for further research in the area of other domains and various types of item formats was pointed out. Also the need for further work in developing comprehensive models which provide the designer or instructor with easy methods for concurrently considering various test related components has been identified.

Reid, J. B. A Monte Carlo comparison of phi and kappa as measures of criterion-referenced reliability. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

The purpose of this study was to compare the values of phi and kappa under several conditions of test score standard deviations,

standard errors of measurement and cutting scores. It was hypothesized that none of the conditions would produce differences in the marginal proportions large enough to cause differences in the corresponding values of phi and kappa.

A random sample generator was used to simulate test scores on a test-retest paradigm under various conditions of population score standard deviation and standard error of measurement.

For each of the 270 sample simulations, the values for both phi and kappa were calculated based on their respective 2 x 2 contingency table. (One dimension was first test score - master-nonmaster - and the other was retest score).

Results - Of the 270 cases simulated, 30 phi-kappa pairs were indeterminate. Of the remaining 240 cases, none showed differences in the first decimal position. Approximately two percent had differences in the second decimal position. About two-thirds of the cases were different in the third and fourth decimal position. Reid concludes that not only are the differences very small, but almost all of those that do occur are of little practical importance. Hence, he suggests that since phi is more widely known and is, in fact, a correlational procedure, the use of kappa should be dropped in favor of phi.

Reid says the advocates of kappa argue:

1) There is potential restriction of variance in CRT's;
2) Therefore classical measurement procedures which rely on variance are not appropriate;
3) as a result, a new index--kappa--is proposed. The implicit assumption is that kappa will react differently to range restrictions than more conventional indices such as phi.

Rentz, R. R. & Rentz, C. C. Does the Rasch model really work? A discussion for practitioners. NCME measurement in education, 1979, 10, No. 2, 1-11.

In this paper, the authors have tried to synthesize the literature related to applications of the Rasch model as it relates to test-development activities. Based on their evaluation of the research literature and their own experiences using the model, they believe that the test developer can feel comfortable in using the Rasch model for constructing tests.

This paper is organized around the following stages in the test-construction or test-development process:

1) defining the content of the test and writing the test items;
2) item analysis; and
3) calibrating the test.

For each stage the authors deal with the current state of knowledge as they see it, offer some recommendations from the literature and their own experience, point out areas where clarification is needed, and call attention to areas of controversy.

Rippey, R. Probabilistic testing. Journal of Educational Measurement, 1968, 5, 211-215.

A computer program was developed for scoring probabilistic tests and different tests were administered to different groups of subjects on four occasions. Automatic increases in reliability were not found. Stereotypical student responses were observed. Non probabilistic multiple choice test scores correlated more highly with essay tests scores than did the probabilistic items, suggesting that a probabilistic scored item did indeed contain a different kind of information than did the non-probabilistic scored items. This difference was probably due to the fact that the probabilistic items contained information which was related to how sure the student was of his knowledge.

Probabilistic tests ask students to assign weights indicating preference or degree of belief for each of the options on a multiple choice item which may or may not have unique correct response options.

Shuford, Albert & Massengill (Psychometrika, 1966, 31, 125-45) have presented a theory of valid confidence testing incorporating the three scoring functions: Logarithmic, Spherical and Euclidean. Rippey's initial purpose was to test the assertion of Shuford, Albert & Massengill that increases in reliability as a result of probabilistic administration and scoring could be anticipated. Automatic increases in reliability were not found.

Rudner, L. M. & Convey, J. J. An evaluation of select approaches for biased item identification. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

The purpose of this study was to investigate the following four approaches to biased item identification using common sets of actual item response data:

1) Transformed item difficulties, which examines the interaction of items and groups;

2) Chi-Square, which determines whether examinees of the same ability level have the same probability of a correct response regardless of cultural affiliation;

3) Item characteristic curve theory (icc) in which differences in the probabilities of a correct response given examinees of the same underlying ability and in different culture groups are evaluated;

4) Factor Score in which item bias is investigated in terms of loadings on biased test factors.

The investigation addressed two questions:

1) Do the four approaches provide identical classifications of items as to their degree of aberrance when applied to item response data corresponding to two culturally different populations?

2) Do the select approaches provide classifications of minimal bias when applied to subsamples of a single population?

Method - The 1973 Stanford Achievement Test (Reading Comprehension subtest), which item for item is deemed equivalent to the hearing impaired version SAT Reading Comprehension Subtest, formed the item pool.

Item responses made by large samples from two diverse culture groups were used in the study.

1) Students in programs for the hearing impaired across the U.S. - this culture group was divided randomly into two subgroups.

2) Students from a large west coast public school system. A major difference between these two culture groups is their exposure to, and ability to use, the English language.

The degree of bias for each item within the SAT was identified by applying a select approach within the transformed item difficulties, icc theory, factor score and chi-square categories to item responses made by:

1) the two diverse culture group samples;

2) two equal culture group samples.

Results - There was some agreement in terms of the identified degrees of aberrance between

1) the transformed item difficulties and chi-square (magnitude) approaches;

2) the icc theory and chi-square (magnitude) approaches; One minus the probabilities associated with the χ^2 and the factor score approach showed little agreement with any of the other methodologies.

Rudner, L. M., Getson, P. R. & Knight, D. L. The effect of various test and item properties on five approaches to biased item detection. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, 1979.

The authors note that the use of certain diagnostic and achievement tests has been viewed recently as contradictory to a basic civil right - the fair treatment of individuals regardless of race, sex, religion or national origin. By identifying and removing biased items in test development, the authors believe that this right can be upheld. Various item bias detection models have been proposed; however, their behavior in practical applications has not been fully evaluated. By using Monte Carlo generated item response data, this research determined the effectiveness, sufficiency and similarity of select biased item detection techniques. Experimental variables included:

1) different proportions of bias for items in the item pool

2) different types of generated item bias

3) different test lengths.

As a result of this research, recommendations are made regarding situations for which each approach is most appropriate.

In conclusion, three of the five investigated techniques--item characteristic curve theory, chi-square with 5 intervals, and transformed item difficulties--produced fairly accurate estimates of the generated amounts of item bias. None of the techniques were found to be ideal. The item characteristic curve theory technique depends on the attainment of accurate parameter estimates, the chi-square technique depends on crude interval estimates of ability, and the transformed item difficulties technique is insensitive to bias in item discrimination.

Samejima, F. Constant information model: A new, promising item characteristic function. Research report 79-1. Personnel and Training Research Programs, Psychological Sciences Division, Office of Naval Research, February, 1979.

The author notes that the methods and approaches thus far introduced for estimating the operating characteristics of item response categories require the Old Test, or a set of items, whose operating characteristics are known. To generalize these methods to apply for the situation where one starts to develop a new item pool, i.e., there is no "Old Test," an approach is made by assuming that the tentative item pool has a substantial number of equivalent items, even though their common item characteristic function is not known yet. It is observed that, within the type of item characteristic function which is strictly increasing in the latent trait θ with zero and unity as its two asymptotes, the area under the square root of the item information function is a constant value, π. The item characteristic function which provides a constant item information is searched and discovered, and is named the constant information model. Using this model, it is observed that the subset of equivalent binary items can be used as a substitute for the Old Test, and those methods and approaches are generalized in the present situation. It is discovered that-for once, items with low discrimination power have a significant role.

Shaycoft, M. F. A paradox in setting cutting scores on criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, San Francosco, 1976.

Three questions are considered:

1) Should the cutting point on the test correspond exactly to the standard set for mastery of the entire domain?

2) If not, should the cutting point be set at the same level for everybody (e.g. for students who have and have not taken a course or studied a module)?

3) If the cutting point is not to be set to correspond exactly to the established standard of mastery, where should it be set, and why?

The binomial distribution provides the basic mathematical model. In the case of multiple choice tests, it is assumed that a binomial distribution governs the probability that an individual will guess a specified number of items correctly, from among all those to which he doesn't know the answer.

No empirical data were used; it was based entirely on theoretical formulas. One table presents the precentage distribution of classification categories, at various stages with both completion and multiple choice tests and various cutting scores. Another table presents the relative amount of misclassification at various stages with various cutting scores.

Results and Conclusions

The optimul cutting point does not generally coincide with the standard set for mastery of the domain. The more competent the group as a whole, the lower the cutting score may be set. These somewhat paradoxical results apply even when the test is of the completion type, so that chance success is not an important factor.

It is noted that to minimize misclassification, the cutting score should be set higher when the test is administered before instruction

than when it is administered after instruction. One practical implication is that when requirements are met by examination only, as a substitute for course attendance plus examination, perhaps it is desirable to set the passing mark somewhat higher.

Shelley, M. F. & Van Mondrans, A. P. The statistical equivalence of items generated from same item form. Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, 1978.

The multiple-choice test items used in this study were computer generated from a simple item form by selecting one correct answer and several incorrect options for each item stem. The results of this study suggest that the utility of generating items by randomly selecting options for a given stem is questionable from a practical as well as a measurement point of view. For example, about 32% more students answered one item stem correctly, when the best distractor was randomly omitted. The data also indicate that the sequence of multiple-choice options has little if any effect on the item statistics.

The authors mention three general techniques of item generation:

1) stores an item skeleton or algorithm and uses random techniques to generate complete items.

2) generates multiple choice questions by selecting from a list of candidates one correct answer and several distractors for a fixed item stem.

3) varies only the sequence of options for a multiple choice item.

In this study, the computer-generated items <u>were only multiple choice items</u>. The computer selected one correct answer and several incorrect options for each item stem.

The results of this study suggest that it may <u>not</u> be desirable to generate multiple-choice items by randomly ordering the options and/or randomly selecting options from a set of candidates. Randomly ordering the same set of options had no discernible effect on the item statistics, did not enlarge the effective size of the item pool, and probably would not be challenging more than once for the same student. Also more computer time is required to randomly sequence options than to randomly order complete items.

This study suggests that the utility of generating items by randomly selecting options for a given stem is questionable on the basis of measurement characteristics, computer programming time, computer execution time, and item writing time.

None of these results apply to completion items.

Shoemaker, D. M. Standard errors of estimate in item-examinee sampling as a function of test reliability, variation in item difficulty indices and degree of skewness in the normative distribution. <u>Educational and Psychological Measurement</u>, 1972, <u>32</u>, 705-714.

Some procedural guidelines are available to aid the researcher in determining the most appropriate number of subtests, number of items per subtest, and number of examinees per subtest. Conspicuous by its

absence in a series of investigations was a systematic examination of the effect on standard errors of estimate due to variations in test reliability. The investigation described in this article was primarily designed to remedy this situation. Additional parameters considered were the variance of item difficulty indices σ_p^2 and degree of skewness in the normative distribution. The parameters estimated were the mean test score μ and the standard deviation of test scores σ.

Shoemaker, J. A study of minimum competency testing programs. National Institute of Education, Office of Testing, Assessment and Evaluation, 1978.

The purpose of this study is to assess, inform and evaluate - not to advocate - with the goal of providing an open, objective and fair description of the minimum competency testing movement. There are two phases of the study. The purpose of Phase I is to collect and disseminate information about minimum competency testing programs. The chief purpose of Phase II will be to implement an evaluation of minimum competency testing programs.

The design of Phase I includes:

1) developing program descriptions;
2) developing program typology;
3) producing program resource guides;
4) designing an evaluation plan.

The design of Phase II includes plans to involve a sample of state and local minimum competency testing programs. The goal will be to assess the representative impact of these programs in a way that will provide needed information about the results of minimum competency efforts. The evaluation will be conducted over a three-year period. It is believed that the outcomes of the study will be valuable to all educational decision-makers who must confront policy issues related to minimum competency testing.

Shoemaker, D. M. & Osburn, H. G. A simulation model for achievement testing. Educational and Psychological Measurement, 1970, 30, 267-272.

A model was developed by the authors that simulates the administration of a single test item to a single examinee. The result is a simulation model of great flexibility for the sampling of both items and individuals.

To obtain type-12 sampling (Lord, F. M. Psychometrika, 1955, 20, 193-200)(random items, subjects & occasions) three features were simulated by the model.

1) The test items: a set of k items must be selected from an item population. Each item must have a difficulty level and a content reference.

2) The examinee: a person with a specified ability level must be randomly selected from a population of people in which the ability under consideration is normally distributed.

3) Testing of examinees over items: does the invidual pass or fail each item in the test?

The model assumes a normally distributed standardized latent ability continuum. The probability of an examinee answering an item correctly is a normal ogive function of his ability level.

The authors employed their simulation model to empirically study certain estimators, the gamma and gamma-stratified coefficients, of test reliability.

Shuford, E. H. Jr., Albert, A. & Massengill, H. E. Admissible probability measurement procedures, Psychometrika, 1966, 31, 125-145.

The authors note that admissible probability measurement procedures utilize scoring systems with a very special property that guarantees that any student, at whatever level of knowledge or skill, can maximize his expected score if and only if he honestly reflects his degree-of-belief probabilities. Section 1 introduces the notion of a scoring system with the reproducing property and derives the necessary and sufficient condition for the case of a test item with just two possible answers. A method is given for generating a virtually inexhaustible number of scoring systems both symmetric and asymmetric, with the reproducing property. A negative result concerning the existence of a certain sub-class of reproducing scoring systems for the case of more than two possible answers is obtained. Whereas Section 1 of this paper is concerned with those instances in which the possible answers to a query are stated in the test itself, Section 2 is concerned with those instances in which the student himself must provide the possible answer(s). In this case, it is shown that a certain minor modification of a scoring system with the reproducing property yields the desired admissible probability measurement procedure.

Sirotnik, K. An analysis of variance framework for matrix sampling. Educational and Psychological Measurement, 1970, 30, 891-908.

Following a brief discussion of the methodology of matrix sampling, this paper attempts to demonstrate the following points:

1) Matrix sampling can be viewed as a simple two factor, random model analysis of variance design, the matrix sampling formulas for estimating the mean and variance being simply the point estimate formulas for estimating components of the underlying linear model.

2) These formulas can be based on the weakest possible set of assumptions,viz., random and independent sampling of examinees and items. No assumptions about the statistical nature of the data need be made.

3) The literature is unclear with respect to the effect of the above sampling assumptions on multiple matrix sampling in the estimation of the mean and especially the variance.

4) Of the three alternative procedures suggested to deal with negative variance estimates in multiple matrix sampling-equating the negative estimates to zero, Winsorizing the distribution of estimates, or treating all estimates alike regardless of sign-the third procedure appears to be the most promising. A simulation study is necessary to determine the shape of the small sampling distribution of variance

components for matrix sampling as well as the relative efficiency of the three methods for handling negative estimates.

Sirotnik, K. Estimates of coefficient alpha for finite populations of items. Educational and Psychological Measurement, 1972, 32, 129-136.

This paper attempts to investigate implications for finite and known item populations of classical test theory and the alpha coefficient among items in paper-and-pencil testing. Finite sampling formulas for α are derived and conceptual problems relating to the treatment of the examinee-item populations are discussed.

The following was shown:

1) An exact estimate of α is possible only in the infinite case; the estimate of α in the infinite case is bounded below and above.

2) If error of measurement variance is conceptualized only as examinee-item response variability, it can not be exactly estimated in either the finite or infinite case. It can be overestimated by MS_{EI}.

3) If error of measurement variance is conceptualized as a residual variance obtained by pooling error and interaction components, it can be exactly estimated only in the infinite case by MS_{EI}. The exact estimate of error of measurement variance conceptualized in this way in the finite case is bounded below and above by $(1 - (m/M))MS_{EI}$ and MS_{EI} respectively.

Sirotnik, K. & Wellington, R. Scrambling content in achievement testing: An application of multiple matrix sampling in experimental design. Journal of Educational Measurement, 1974, 11, 179-188,

This study is designed to research the question of scrambling item content in the construction of achievement tests, in order that general implications could be drawn for both examinee and item populations. To achieve this generality, the methodology of multiple matrix sampling was combined with a simple two-group experimental design: a random group of eighth graders responded to mathematics, science, social studies, reading and language arts achievement items organized in a scrambled (random) test format, while another random group responded to the same items organized in a fixed (segregated by subject matter) test format. The results indicate that scrambling cognitive test items has minimal or no effect on mean examinee test performance or on any of the other parameters included in the analysis.

Skager, R. W. The great criterion-referenced test myth. CSE Report No. 95, Los Angeles: Center for the Study of Evaluation, 1978.

In this paper, Skager reviews the history of criterion-referenced measurement which began when Glaser separated tests into those that are norm-referenced (NRT) and those that are criterion-referenced (CRT) as a way of emphasizing the distinction between scores that can be interpreted in terms of what a person actually can do vs. how well a person

does compared to other people. The author emphasizes that it is the interpertation of tests and the way in which test content is specified and not the test itself which can be criterion- or norm-referenced.

Skager notes that whether or not the test measures an open or a specified content domain is a direct function of the ways in which their content domains are specified. Taking the function for which the test is to be used as the primary dimension, the author says that it turns out that certain content specification modes are more appropriate for certain functions than others and that particular types of score interpretations go with particular function/specification mode combinations. He then attempts to show how this is the case, limiting the discussion to the use of educational tests in the evaluation of learners and not in the evaluation of the conditions under which learning occurs. The two groups of functions for tests in the evaluation of learners which are referred to as "formative" and "summative" after Bloom, Hastings and Madaus '71 are explained.

Modes for specifying test content which fall into the following five categories are discussed:

a) the content/process matrix;
b) the theoretical construct;
c) criterion sampling;
d) objectives based;
e) formal item generation rules.

Three types of scoring interpretations which involve referencing an examinee's test performance to

a) a content domain;
b) a criterion or standard;
c) relative position in some defined reference population of persons are discussed.

This paper makes the point that not one type of test is solely criterion-referenced as compared to another type of test that is solely norm-referenced. The really important differentiating factors have to do with the function for which the test is to be used and the mode by which the test content is to be specified. Once this distinction is made, criterion- vs. norm-referencing becomes a matter of the type of score interpretation which is likely to be more useful. In many situations, both types of interpretations are likely to be useful, regardless of the mode of content specification.

Smith, D. U. The effects of various item selection methods on the classification accuracy and classification consistency of criterion-referenced instruments. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

This study examined the effects of certain item selection methods on the classification accuracy and classification consistency of criterion-referenced instruments. Three item response data sets, representing varying situations of instructional effectiveness, were simulated. Five methods of item selection were then applied to each data set for the purpose of selecting a best subset of items. These were:

1) RPB - point biserial correlation between the dichotomous item response, (pass, fail) and the total test score.

2) CV - the difference between the posttest and pre-test proportion of examinees answering the item correctly.

3) BR - the difference between the proportion of examinees in the mastery group and the nonmastery group passing the item.

4) PHI - the phi coefficient computed between the dichotomous item response versus mastery/nonmastery status on the total test.

5) RDM - a random selection of the items.

Parallel forms of an instrument were then constructed and two measures of test quality obtained: the proportion of correct classifications (for classification accuracy) and a mean split-half coefficient of agreement, coefficient beta, (for classification consistency).

A Kruskal-Wallis one-way anova of ranked data was performed on the proportion of correct classifications for each of the five methods of item selection. Pairwise comparisons were then performed on the mean rankings for each of the five selection methods. A similar approach to the analysis of classification consistency was taken.

Results indicated that methods yielding the best accuracy and consistency varied across the situations of instructional effectiveness.

Results of the analysis of classification accuracy and classification consistency revealed that no single method of item selection resulted in instruments which were consistently superior to other methods in correctly classifying examinees when instructional effectiveness was varied.

Overall, the findings reported in this paper suggest that the classification accuracy and classification consistency of CR assessment instruments can be improved through the use of item selection methods. When a "fair amount" of variability is present in the distribution of test scores, the RPB method was superior to the other methods. However, in true mastery-learning situations, the CV method of item selection is recommended. Finally, in those situations where instruction is relatively ineffective, the BR and PHI methods tended to yield the best instruments.

Stallings, W. M. & Gillmore, G. M. A note on "accuracy" and "precision." Journal of Educational Measurement, 1971, 8, 127-129.

In the literature of engineering and "hard" sciences, the term precision shares a common core meaning with reliability as used by behavioral scientists. Accuracy and validity have a similar semantic overlap.

In educational and psychological measurement, there is an interchangeable usage of accuracy and precision in defining reliability.

The authors of this paper advocate the use of precision, rather than accuracy, in describing reliability.

Subkoviak, M. J. Empirical investigation of procedures for estimating reliability for mastery tests. Draft of paper submitted for publication.

Four different procedures (Huynh, '76; Marshall & Haertel, '76; Subkoviak, '76; Swaminanthan, Hambleton and Algina, '74) have been

proposed for estimating the proportion of persons consistently classified as master/master or nonmaster/nonmaster on two mastery tests. Estimates of this proportion were obtained for repeated samples of size N=30 for each of the above procedures. The estimates were then compared for accuracy to the value of this proportion in the population on N=1586 subjects from which the samples were drawn. Both test length and mastery criterion were varied. While reasonably accurate estimates were generally obtained for all four procedures, instances of systematic estimation bias were observed.

All four procedures (Huynh, '76; Marshall & Haertel, '76; Subkoviak, '76' Swaminathan et al., '74) appear to provide reasonably accurate estimates of the proportion of consistent classifications on two mastery tests, for the various cases considered. However, relative advantages and disadvantages can be noted. The Swaminathan procedure produces unbiased estimates; but it requires two testings and standard errors are relatively large for classroom size samples. On the other hand, the Hyynh, Marshall-Haertel, and Subkoviak approaches require only one testing; and standard errors of estimate are relatively small; but for short tests, each procedure appears prone to a different type of systematic bias. All things considered, the Huynh approach seems worthy of recommendation. It is mathematically sound, requires only one testing and produced reasonably accurate estimates, which appear to be slightly conservative for short tests.

Subkoviak, M. J. Estimating reliability from a single administration of a mastery test. Journal of Educational Measurement, 1976, 13, 265-276.

The author points out that a number of different reliability coefficiants recently have been proposed for tests used to differentiate between groups such as masters and non-masters. One promising index is the proportion of students in a class that are consistently assigned to the same mastery group across two testings. The present paper proposes a single test administration method for estimating this index. This is achieved by substituting assumptions for the second test administration. These assumptions are:

1) parallel test scores X_i and X_i are independently distributed as
2) identical binomial distributions for a fixed person i.

Subkoviak, M. J. The reliability of mastery classification decisions. University of Wisconsin unpublished paper, 1979.

The author notes that Hambleton, Swaminathan, Algina and Coulson distinguish three different concepts of reliability that arise in the context of criterion-referenced testing:

a) reliability of mastery classification decisions;
b) reliability of criterion-referenced test scores;
c) reliability of domain score estimates.

This paper focusses on the first of these three types of reliability.

Specifically, the paper is concerned with approaches to criterion-referenced reliability that emphasize the consistency of mastery and non-mastery decisions over repeated testing of the same group.

Four methods of estimating reliability of mastery classification decision are outlined:

1) Carver method;
2) Swaminathan-Hambleton-Algina method;
3) Huynh method;
4) Subkoviak method.

A comparative analysis of the four reliability methods providing some insight into the various strengths and weaknesses of the procedures is presented.

The reliability of mastery classifications versus the reliability of mastery test scores is also discussed.

Subkoviak, M. J. & Wilcox, R. R. Estimating the probability of correct classification in mastery testing. Paper presented at the annual meeting of the American Educational Research Associaton, Toronto, 1978.

A procedure is proposed for estimating the proportion of persons in a group that are correctly classified on a mastery test, i.e., the proportion whose observed classification agrees with their true classification. A numerical example is provided, and expansions of the procedure are discussed.

Following Keats and Lord (1962), the true score for an individual is defined as the mean of an individual's observed proportion correct scores over repeated parallel tests and is assumed to have as distribution for the population of examinees some member of the Beta family of distributions. For predetermined mastery criterion expressed as the number of items correct on a fixed length mastery test, a procedure is proposed for estimating the proportion of persons in a group that are correctly classified as master and nonmaster, i.e. the proportion whose observed classification agrees with their true classification. This is the probability of correctly classifying any given individual in the group. It is suggested that this probability would tend to increase as the density of true scores about the mastery criterion decrease and as the number of items on the test increases. A numerical example is given and an extension to classification of three or more levels of mastery is discussed.

Swaminathan, H. & Gifford, J. A. Estimation of parameters in the three-parameter latent trait model. Paper presented at an AERA-NCME sym posium entitled "Explorations of Latent Trait Models as a Means of Solving Practical Measurement Problems", San Francisco, 1979.

The purpose of this study was to compare two methods for estimation of parameters in the three-parameter logistic model, the Urry method of estimation and the maximum likelihood procedure. The computer programs that were used were the ANCILLES and the LOGIST. The efficiency of the

procedures were compared with respect to the accuracy of estimation, the effect of violating underlying assumptions (for the Urry procedure), and the statistical properties of the estimators. The factors that were controlled were: test length (4 levels), examinee population size (3 levels) and ability distribution (3 levels).

The results indicate that, in general, the maximum likelihood procedure is superior to the Urry procedure with respect to the estimation of all item and ability parameters.

The number of examinees had a slight effect in improving the accuracy of estimation of the difficulty, and the chance-level and ability parameters. However, increasing the number of items and the number of examinees considerably improved the accuracy of the discrimination estimates with both procedures.

Although the maximum likelihood estimates were superior to the Urry estimates, especially in the case of short tests, the difference between them was negligible when the number of items and the number of examinees increased. This is of importance, since the Urry procedure requires considerably less computer time than the maximum likelihood procedure. However, the Urry procedure, in general, deletes more items and examinees during the estimation than the maximum likelihood procedure. This may explain the rapidity of convergence and indicate a weakness in the Urry procedure.

The bias and consistency results indicate that for small numbers of items, the estimates of the item and ability parameters are biased, with the Urry estimates being more biased than the maximum likelihood estimates. As the number of examinees and the number of items increase, it appears that the estimators are unbiased, and in fact, are consistent. This in a sense supports a conjecture of Lord's and shows that the three-parameter model may be statistically viable.

Swaminathan, H. Hambleton, R. K. & Algina, J. A Bayesian decision-theoretic procedure for use with criterion-referenced tests. Journal of Educational Measurement, 1975, 12, 87-98.

The authors present an exposition of a decision-theoretic solution to the problem of allocating individuals to mastery states on the objectives included in a criterion-referenced test. Decisions are made by taking into account prior and collateral information on the examinees and also the losses associated with misclassifications.

Swaminathan, H., Hambleton, R. K. & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-267.

The purpose of this article was to describe a decision-theoretic formulation of criterion-referenced test reliability. It has been suggested that the primary purpose of criterion-referenced testing in objective-based instructional programs is to classify examinees into mastery states (masters or non-masters) on the objectives included in the test. The authors define the reliability of CR test scores in terms of consistency of the decision-making process across repeated administrations of the test. Specifically, reliability of a CR test is defined

as a measure of agreement above chance expectation between the decisions made about examinee mastery states in repeated test administrations for each objective measured by the criterion-referenced test.

Coefficient Kappa (k) takes into account the measure of agreement expected by chance alone. $k = (p_o - p_c)|(1 - p_c)$ where p_o is the observed proportion of agreement and p_c is the expected proportion of agreement. k has an upper limit of +1 and lower limit of close to -1. Since we usually have only a sample of examinees, k must be estimated. $\hat{k}$ is defined as the sample analogue of $k = (p_o - p_c)|(1 - p_c)$.

The authors conclude that the coefficient of agreement k and hence the reliability of CR subtests is dependent on factors that affect the decision process. These factors include:

1) the method of assigning examinees to mastery states,
2) selection of the cutting score,
3) test length,
4) heterogeneity of the group.

The authors state that decision-making consistency is a measure of the reliability of the entire decision-making process, and that the test itself is only one input into the decision-making process. In generalizing reliability data to a new decision-making situation, all factors that affect the process must be considered.

Swinton, S. S. On the reliability of item clusters. Paper presented at the National Council on Measurement in Education, San Francisco, 1979.

A result of Henrysson is used to obtain more accurate estimation of reliability than is offered by the commonly used KR-20 approximation. Applying this result to examples and to data from a statewide assessment program reveals that the KR-20 formula may and frequently does seriously underestimate these reliabilities. Implications for interpretation of criterion-referenced tests and for the relationships of single-item reliability estimates to test reliability estimates are discussed.

Tatsuoka, K. Final report: Analytical test theory model for time and score. CERL Report E-8, Computer-based Education Research Laboratory, University of Illinois, Urbana, Illinois, July, 1979.

The purpose of this study was to find a way to utilize response-time data in the scoring procedure of achievement testing. The empirical study of adaptive diagnostic testing and a computerized instructional system revealed that the differences in type of information-processing skill developed by different instructional backgrounds affect the learning of further instructional materials to a great extent. The author urges us to tap what information-processing strategy was used to respond to a given problem, not only considering individual differences in ability or achievement level, derived solely from the performance scores. It was seen that there were response patterns which yielded the same achievement level θ but the answers were obtained by two different approaches. If these individual differences in information-processing skills would not be detected on a diagnostic test, then a proactive

inhibition effect will cause a serious learning deficiency, even for many good students, upon studying further instructional materials.

A model useful in identifying discriminating items that are sensitive to differences in instructional method was developed. It also is helpful in identifying an individuals' instructional background to a certain extent.

A method that estimates θ by regressing θ onto test-items was developed and compared with θ estimates by the maximum likelihood method. The new method provided as good θ as the traditionally established method did. This method is powerful when the number of subjects and test-items available are small. Also estimates are always obtainable and moreover free from a choice of ordering test-items.

Tatsuoka, K. The least-squares estimation of latent trait variables by a Hilbert space approach. Computer-based Education Research Laboratory Report E-4, February, 1979.

This research developed a new method for estimating a given latent trait variable θ by the least-squares approach. The notion of multiple regression equation was reinterpreted in terms of properties of a Hilbert space and the calculation formula for beta weights that can be obtained recursively in the form of Fourier series was derived. The θ values estimated by this method and the maximum likelihood method were compared using live data.

Tatsuoka, K., & Birenbaum, M. The danger of relying solely on diagnostic adaptive testing when prior and subsequent instructional methods are different. CERL Report E-5, Computer-based Education Research Laboratory, University of Illinois, Urbana, Illinois, March, 1979.

A computerized diagnostic adaptive test for a series of pre-algebra signed-number lesson was programmed along with a computer-managed routing system by which each examinee was sent to the instructional unit corresponding to the level of skill at which she/he stopped in the initial test. Upon completion of the course a computerized conventional post-test was given to the examinees. The posttest scores were not unidimensional, while the pretest and post-test data obtained from a previous study indicated a strong tendency to be unidimensional. The response patterns of the posttest in the present study showed a high error rate for the skills prior to stopping levels for a subgroup of examinees.

A cluster analysis was performed on the response patterns and four different groups were found. A discriminant analysis indicated significant differences among the four groups in response patterns of the skills in signed number operations. After interviewing the teachers and several students, the authors concluded that there was a proactive inhibition effect.

The scoring procedure of the adaptive testing did not consider individual differences in information-processing skills which were affected by the instructional method used in previous teaching. Thus, the students who were taught to perform the beginning part of a set of

hierarchically ordered skills by instructional method A would very likely get confused in a lesson in which a different instructional method B was adopted. Consequently, quite a few number of peculiar response patterns were seen in the performance on the posttest. This fact led the authors to caution that one should be careful not to rely solely on test results determined by performance scores on a diagnostic pretest when a computer-managed instructional system is to route each examinee to their level of instruction.

Tatsuoka, K. & Tatsuoka, M. A model for incorporating response-time data in scoring achievement tests. CERL Report E-7, Computer-based Education Research Laboratory, University of Illinois, Urbana, Illinois, July, 1979.

The differences in type of information-processing skill developed by different instructional backgrounds affect, negatively or positively, the learning of further advanced instructional materials. That is, if prior and subsequent instructional methods are different, a proactive inhibition effect produces low achievement scores on a posttest. This fact poses a serious problem for routing of students to an instructional level on the sole basis of performance on a diagnostic adaptive test.

The authors state that it is essential that what information-processing strategy was used be unraveled and that this knowledge be considered simultaneously.

They note that response time often provides supplementary information which differentiates among individuals showing identical quality of performance. A model that reflects this kind of information, obtainable from response time scores, is formulated in a similar manner to latent trait theory and is discussed. This model is useful in identifying discriminating items that are sensitive to differences in instructional method. It also is helpful in identifying an individual's instructional background to a certain extent.

Toothaker, L. E. & Chang, H. S. Analysis of qualitative data: A comparison of various methodologies. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

This research empirically compared several methods of analyzing qualitative data for a variety of settings. The results showed that the Pearson chi-square test is to be recommended for the situations examined in this research: equal group size, ten observations per call (average), designs ranging from 2x2 to 5x4, and 2x2x2 to 3x3x2, and various probability patterns. The ANOVA F-test, CATANOVA (an analysis of variance for categorical data), and logit chi-square all have some limitations and would be second choices to the Pearson chi-square test.

Tucker, S. B. & Vargas, J. S. Item analysis of criterion-referenced tests for a large individualized course. Unpublished paper.

For item analyzing criterion-referenced tests, the authors suggested using the difference in difficulty indices between pre-and

posttests, which they called the Pre-post Difference Index (PPDI). This paper described this use and investigates in particular the sensitivity of the PPDI to changes in instruction. Tests were given as pre- and posttests for each of 10 self-instructional modules which the students go through in a learning center. The students were approximately 800 education majors. In addition to revealing weaknesses in the items, the PPDI revealed weakness in the instruction.

Conclusion: Like traditional item analysis procedures for norm-referenced tests, the authors have found the PPDI to be useful in identifying weak items in criterion-referenced tests. In individualized instruction, one has certain criteria one wishes students to meet, and they are operationalized in the posttests. There is a limit, therefore to how much one would wish to improve an instructional system by re-writing items. Any item analysis index for criterion-referenced tests should therefore help the instructor improve student learning as well as his items. Improvement in instruction should be reflected in the item analysis data.

Although the results of the present study were not highly significant statistically, they gave evidence that the PPDI is sensitive to as small a change in instruction as a 10 page program. Although it is not based on highly sophisticated theories of test construction and analysis, the PPDI has been found to be useful in flagging not only weak items, but weaknesses in instruction.

Unruh, R. P. Test Validity: Process and product dimensions. Paper presented at the annual conference of the California Society of Educational Program Auditors and Evaluators, San Francisco, 1979.

The author states that variables which operate a typical learning situation may be classified into two broad categories: product variables: (i.e., what is learned, or content) and process variables (i.e., how the content is learned). In achievement testing careful attention has been given to product variables, however, process variables have generally been overlooked. The APA Standards for Educational and Psychological Tests recognizes three major types of validity: criterion-related, content, and construct. Each of three types of validity deal with some aspect of expected/observed outcome (product) variables. At the same time, the Standards contains no recommendations regarding that class of variables which might be called process or process-related. Because individuals exhibit different learning styles, and because different educational programs feature different learning environments and styles (e.g., rote vs. discovery, individual vs. group, etc.) it is important that individuals who use achievement tests make allowances for the effects of learning processes on the products of learning. For these reasons it is suggested that users of tests should carefully examine the relationship between the processes involved in the learning and the testing situation as well as the relationships between the products. Several variables which may be important when assessing the process validity of a particular test for a program are suggested.

Walker, C. B. Control test items: A baseline measure for evaluating achievement. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

The author points out that the evaluator gives a posttest of achievement to pupils in a given program, then tries to interpret the scores by comparing them with some type of performance baseline, norm or expectancy. Two common types of baselines are provided by time series data and by independent control groups. In this paper another type of comparison is discussed for which the name control test items (or control items) is coined. The following topics are discussed: rationale and precedents for the method, advantages and disadvantages, sources of possible control objectives and unresolved issues.

The method proposed here consists of giving one group of program pupils a test of two comparable sets of skills, namely program and control objectives. The familiar control group baseline as well as the control items baseline are estimates of how program pupils would have performed on program objectives without the benefit of relevant instruction. In the traditional case one draws inferences from one group of pupils to another, but in the case of control items the inference is from one group of skills to another. Control test items are thus a conceptual analogue of control groups.

The author lists three advantages of control items:

1) They give another indicator of program effects that can be used along with time series and control groups.

2) Control test items provide a method of comparison to use when time series cannot be used and when control pupils are completely unavailable, available in insufficient numbers, or not clearly comparable to the treatment groups.

3) They are not subject to some of the problems that control groups have. E.G. Since control items give a within-test, within-pupils baseline, treatment and control measures of program pupils can be compared without regard to program leakage.

Three disadvanteges of the control test item method are:

1) They do not give the type of information that control groups do, namely, a comparison between existing programs.

2) It would be harder, although not impossible, to use the control items approach outside of an objectives-based teaching and testing context.

3) The use of control items may reduce the reliability of the total scores on program items.

The author lists the five following sources of possible control objectives:

1) Look for skills which are relatively circumscribed or discrete.

2) Look for skills which used to be taught and learned at the relevant level, but which are not now fashionable.

3) Look for skills which are traditionally taught at higher levels but which are not inherently harder to learn at the level tested.

4) When a program objective samples only a part of a well-defined content area, like irregularly spelled words, control items may be selected from the excluded part of that content.

5) Objectives from outside the content domain of the program are available.

Some issues to be resolved involve those of sampling and formatting.

Werts, C. E., Linn, R. L. & Jöreskog, K. A congeneric model for platonic true scores. Educational and Psychological Measurement, 1973, 33, 311-318.

The authors provide an alternative formulation [to Levy (Psychological Bulletin, 1969, 71, 276-277)] which allows for the model parameters to be determined given the structural specification of zero mean error and independence among errors for different items and between errors and true scores. Their approach is drawn from latent structure analysis (U. Grenander (Ed.), Probability and statistics, the Harold Cramér volume. New York: Wiley, 1959, 9-38) for the special case of dichotomous latent variables.

Werts, C. E., Linn, R. L. & Jöreskog, K. G. Intraclass reliability estimates: Testing structural assumptions. Educational and Psychological Measurement, 1974, 34, 25-33.

Intraclass correlation reliability estimates are based on the assumption that the various measures are equivalent. Jöreskog's (Biometrika, 1970, 57, 239-251) general model for the analysis of covariance structures can be used to test the validity of this assumption.

Whitely, S. E. & Dawis, R. E. The nature of objectivity with the Rasch model. Journal of Educational Measurement, 1974, 11, 163-178.

Rasch and Wright have claimed that the Rasch model leads to a higher degree of objectivity in measurement than has been previously possible. Whitely and Dawis found that this is not so.

The authors conclude that the lack of impact of the Rasch model in test development is due more to the current status of trait measurement than to the properties of the model. Many of the advantages of the Rasch model necessitate a different kind of data for trait measurement than is now characteristic of the field. Explicit trait-item theory, locally independent items and routine administration of tests by computer, would be part of the necessary technological sophistication.

Wilcox, R. R. Achievement tests and latent structure models. Center for the Study of Evaluation, University of California at Los Angeles, 1977.

Existing latent structure models intended to describe achievement tests either characterize a population of examinees in terms of a few specific items or they characterize an item domain in terms of a single examinee. For many situations it is desired to do both. This paper describes a possible solution to this problem for certain important special cases. Explicit estimates of the parameters of the model are given. The present model generalizes the models described by Wilcox and Morrison and it contains the binomial error model as a special case.

Wilcox, R. R. Analyzing the distractors of multiple choice test items. Center for the Study of Evaluation, University of California at Los Angeles, 1979.

Wilcox states that when analyzing the distractors of multiple-choice test items, it is sometimes desired to determine which of the distractors has a small probability of being chosen by a typical examinee. At present, this problem is handled in an informal manner. In particular, using an arbitrary number of examinees, the probabilities associated with the distractors are estimated and then sorted according to whether the estimated values are above or below a known constant p_0. In this paper a more formal framework for solving this problem is described. The first portion of the paper considers the problem from the point of view of designing an experiment and a later section considers methods that might be employed in a retrospective study. Brief consideration is also given to how an analysis might proceed when a test item has been altered in some way.

Wilcox, R. R. Determining the length of a criterion-referenced test. First draft of a paper to appear in a special issue of APM. Center for the Study of Evaluation, University of California at Los Angeles, July, 1979.

The first purpose of this paper is to give a brief review and critique of the three general approaches that might be used when determining the length criterion-referenced test. Secondly, new results on test length are described. Finally, possible directions for future research are indicated.

Wilcox states that the most important point of his paper is that there is no magic number or even magic formula for determining test length. Even within the seemingly narrow problem of comparing an examinee's true score to some constant, there are many approaches to the problem. Moreover, in terms of which true score to use, it is not at all clear as to what extent the three types considered here are in competition with one another. For the moment, the author states that the best we can do is to be very precise about what we want to determine, consider what assumptions we are willing to make, and act accordingly.

Wilcox, R. R. On false-positive and false-negative decisions with a mastery test. Center for the Study of Evaluation, University of California at Los Angeles, 1979.

In an earlier paper Wilcox examines two methods of estimating the probability of making a false-positive or false-negative decision with a mastery test. Both procedures make an assumption about the form of the distribution of true scores over the population of examinees which might not give good results in all situations. In this paper upper and lower bounds on the two possible error types are described which make no

assumption about the true score distribution beyond that its first two moments exist. The first method depends only on the ability to determine the mean and variance of the true score distribution. Wilcox indicates that such estimates are readily available when the binomial or compound binomial error model is assumed. The second method is based on the binomial error model which is frequently used to describe a mastery test. Illustrations are given on how these bounds might be used to determine the length of the test.

Wilcox, R. R. Upper and lower bounds to the probability of a false-positive or false-negative decision with a mastery test. Unpublished paper, 1978.

In a previous paper Wilcox described two methods of estimating the probability of a false-positive or false-negative decision with a mastery test. Both procedures make assumptions about the form of the true score distribution which might not give good results in all situations. In this paper, upper and lower bounds on the two possible error types are described which make no assumption about the form of the true score distribution.

The first method depends only on our ability to determine the mean and variance of the true score distribution. Such estimates are readily available when the binomial or compound binomial error model is assumed. The second method is based on the binomial error model which is frequently used to describe a mastery test.

Illustrations are given on how these upper and lower bounds might be used to determine the length of the test.

Wilcox, R. R. & Yeh, J. P. Using latent structure models to measure achievement. Center for the Study of Evaluation, University of California at Los Angeles, 1977.

This paper considers the use of certain types of latent structure models when measuring the achievement of examinees. The first portion of this paper describes how certain exact solutions might be used as approximations to more general solutions or as initial estimates in some iterative estimation technique. They also derive some new exact results for the case of hierarchically related items. Finally, they illustrate that these models provide a correction for guessing which may lead to different results than those obtained using the usual observed scores when comparing two populations of examinees.

It is suggested that there may be technical problems when using iterative estimation schemes for maximum likelihood estimation and that it is often difficult to determine whether these schemes yield reasonable results. For the model described by Macready and Dayton ('77), an approximate alternative is proposed using the exact algebraic solutions for the parameters in the case of separate item pairs. Using two separate 2x2 tables, the estimates are shown to be reasonably close to the solution provided by the iterative procedures involving all four items.

A model and estimation procedure for two hierarchically related achievement test items are discussed. (This model is a compliment of

the item sensitivity model - i.e. by assuming that the probability of guessing is greater than zero, and the probability of incorrect response given that the examinee does not know the correct response is zero). However, it is pointed out that not all the parameters can be estimated simultaneously. To provide estimates for all the parameters, it is suggested that a third appropriately chosen item could be used. A numerical example is given.

When comparing two populations of examinees, it is suggested that by assuming a probabilistic response model, the examiner may arrive at a conclusion opposite to that of using the observed number-correct scores. This result is also generalized to n-items tests.

The hierarchical model is then modified for completion items, i.e., by assuming that the probability of not knowing and guessing the correct response to be zero and the probability of knowing but giving an incorrect response to be greater than zero. (This is the same as the item sensitivity model.)

Finally, it is also suggested that the models could be used to determine whether a multiple-choice item is "ideal" in the sense that the probability of guessing is equal to the recriprocal of the number of choices (Weitzman, 1970).

Woodson, M.I.C.E. Classical test theory and criterion-referenced scales.

The author notes that for criterion-referenced scales the range of interest is defined by a range of the characteristic rather than the distribution of that characteristic in some population. The calibration sample must be representative of that range of interest. When the range of interest is appropriately defined, an appropriate calibration sample may be selected, and classical test theory applies directly to criterion-referenced scales.

Woodson, M.I.C.E. The issue of item and test variance for criterion-referenced tests. Journal of Educational Measurement, 1974, 11.

The author points out that it has been argued that item variance and test variance are not necessary characteristics for criterion-referenced tests, although they are necessary for norm-referenced tests. The author believes that this position is in error because it considers sample statistics as the criteria for evaluating items and tests. Within a particular sample, an item or test may have no variance, but in the population for which the test was designed and evaluated the author believes that both items and tests must have variance.

Woodson, M.I.C.E. The issue of item and test variance for criterion-referenced tests: A reply. Journal of Educational Measurement, 1974, 11.

In his reply to Millman and Popham, the author states that it is a necessary condition that items and tests have variance and discrimination

in the range of interest (population of observations) for which they are calibrated and selected. The basis for selection of the calibration sample determines the kind of scale which will be developed. A random sample from a population of individuals leads to a norm-referenced scale, and a sample representative of abilities of a range of a characteristic leads to a criterion-referenced scale.